

Minería de textos georreferenciados aplicada al contexto de robos en Lima, Perú



viu

**Universidad
Internacional
de Valencia**

Titulación:

Máster Universitario en Big Data
y Ciencia de Datos
Curso académico
2025 – 2026

Alumno:

Fernández Egúsquiza Alexander
Frei
D.N.I: 42453045
Directora de TFM: Huerta Pacheco
Nery Sofía

Convocatoria:

Primera

De:

 Planeta Formación y Universidades

Índice

Resumen	4
Abstract	5
Capítulo 1: Introducción	6
1.1 Antecedentes	6
1.2 Planteamiento del problema	8
1.3 Objetivos	9
1.3.1 General	9
1.3.2 Específicos	9
1.4 Justificación	9
Capítulo 2: Estado del arte	11
2.1 Minería de datos	11
2.2 Minería de texto	17
2.4 Georreferenciación	20
2.5 Técnicas estadísticas para datos textuales	22
2.5.1 Tabla de frecuencias	22
2.5.2 Tabla de contingencias	22
2.5.3 Análisis de correspondencia	22
2.5.4 Análisis cluster	22
2.6 Otras herramientas para analizar datos textuales	23
2.7 Marco de referencia	24
Capítulo 3: Desarrollo del proyecto	26
3.1. Metodología	26
3.2 Descripción de los datos	29
Capítulo 4: Hallazgos encontrados	31
4.1. Extracción de datos	31
4.2. Análisis de los datos	31
4.2.1 Descripción	31
4.2.2 Clasificación	42
Capítulo 5: Conclusiones	48
5.1. Conclusiones del trabajo	48
5.2. Limitaciones técnicas y oportunidades de mejora	48
5.3. Trabajos a futuro	48
Capítulo 6: Referencias	49



Capítulo 7: Anexos.....	52
--------------------------------	-----------

Índice de Cuadros

Cuadro 1.1: Tasa de víctimas de robo o intento de robo (por cada 100 habitantes de 15 años y más) en Lima Metropolitana (enero 2020–diciembre 2024)	5
Cuadro 2.1: Comparativo entre Minería de datos y Minería de textos	14
Cuadro 3.1: Resumen final: Cantidad de noticias por año y por diario (2020–2024)	
.....	28
Cuadro 4.1: Frecuencia de palabras asociadas a Robo - Período 2020 a 2024	29
Cuadro 4.2: Frecuencia de robos por Distritos - Período 2020 a 2024.....	30
Cuadro 4.3: Frecuencia de palabras asociadas a Robo - Período 2020 a 2021	32
Cuadro 4.4: Frecuencia de robos por Distritos - Período 2020 a 2021.....	33

Índice de Figuras

Figura 1.1: Diferencia entre minería de texto y minería de datos.....	6
Figura 1.2: Flujo de trabajo: web scraping, clasificación de texto, generación de indicadores de innovación, almacenamiento y consulta.....	7
Figura 3.1: Etapas de la presentación del conocimiento	27
Figura 4.1: Nube de palabras asociadas a Robo – periodo 2020 al 2024.....	35
Figura 4.2: Nube de palabras asociadas a Robo – periodo 2020 al 2021.....	36
Figura 4.3: Nube de palabras asociadas a Distritos – periodo 2020 al 2024	37
Figura 4.4: Nube de palabras asociadas a Distritos – periodo 2020 al 2021	37
Figura 4.5: Dendogramas de palabras asociadas a Robo – periodo 2020 a 2024	38
Figura 4.6: Dendogramas de palabras asociadas a Robo – periodo 2020 a 2021	39
Figura 4.7: Dendogramas de palabras asociadas a los Distritos – periodo 2020 a 2024	40
Figura 4.8: Dendogramas de palabras asociadas a los Distritos – periodo 2020 a 2021	40



Resumen

El presente trabajo integra minería de textos, técnicas de web scraping y procesos de georreferenciación dentro del marco metodológico del Descubrimiento de Conocimiento en Bases de Datos (KDD), con el propósito de identificar patrones espacio-temporales asociados a noticias sobre robos ocurridos en Lima, Perú, durante el periodo 2020–2024. La metodología permitió estructurar el proceso investigativo en fases secuenciales: selección, preprocesamiento, minería, evaluación y presentación; garantizando coherencia analítica y trazabilidad técnica desde la recolección de datos hasta la generación de conocimiento aplicable.

La base de datos fue construida mediante web scraping aplicado a los diarios: El Comercio, Gestión, Correo y Ojo, utilizando lenguaje R, obteniéndose un total de 32 411 registros estructurados en título, resumen y fecha. Posteriormente, se desarrollaron procesos de limpieza semántica, normalización léxica, eliminación de stopwords y georreferenciación por distrito, asegurando consistencia textual y espacial para el análisis.

En coherencia con el marco metodológico del KDD, la fase de minería de datos y minería de textos constituyó el núcleo analítico del estudio, siendo el momento en el cual se aplicaron algoritmos y técnicas estadísticas para la identificación de patrones significativos. Dentro de esta etapa se emplearon tablas de frecuencia, matrices de contingencia, análisis de correspondencias, análisis cluster jerárquico representado mediante dendrogramas, así como visualizaciones complementarias mediante nubes de palabras y mapas temáticos. Esta implementación permitió transformar los datos textuales previamente depurados y georreferenciados en estructuras analíticas interpretables, asegurando que el descubrimiento de patrones espaciales y temáticos se realizara bajo un flujo metodológico consistente con las fases establecidas por el modelo KDD.

Los resultados obtenidos evidenciaron que los distritos con mayor incidencia relativa de noticias asociadas a robos durante el periodo 2020–2024 fueron: Independencia, La Victoria y Miraflores, mostrando patrones espaciales persistentes y coherentes con la estructura urbana de Lima. Estos hallazgos demuestran la capacidad de la minería de textos georreferenciada para identificar zonas críticas a partir de información periodística digital, consolidando una metodología replicable para el análisis criminológico urbano y el apoyo a la toma de decisiones en seguridad ciudadana, no solo en Lima, sino en otras ciudades del Perú y del Mundo.

Palabras clave: minería de textos, web scraping, georreferenciación, análisis espacial, robos, KDD.



Abstract

The present study integrates text mining, web scraping techniques and georeferencing processes from the Knowledge Discovery in Databases (KDD) methodology in order to identify space-time patterns associated with assaults that occurred in Lima, Peru from 2020 to 2024. The methodology allowed to structure the research process in sequential phases: selection, preprocessing, mining, pattern evaluation and knowledge presentation; guaranteeing analysis coherence and technical traceability from data collection to applicable knowledge generation.

The database was built using web scraping techniques applied to the following newspapers: El Comercio, Gestión, Correo and Ojo, using R language, obtaining a total of 32 411 registries structured in title, summary and data. Afterwards, we developed processes of semantic cleaning, lexical normalization, stop words removing and by district georeferencing, assuring textual and spatial consistency for the analysis.

In accordance with the KDD methodology, the data mining and text mining phases constituted the study analytic core, at the moment in which algorithms and statistical techniques for trend patterns identification were used. In this step we employed frequency tables, contingency matrices, correspondence analysis, hierarchy cluster analysis with dendrograms, as well as complementary visualizations via word clouds and thematic maps. This implementation allowed us to transform previously refined and georeferenced text data into interpretable analytic structures, assuring that the discovery of spatial and thematic patterns would be done under a methodology flow consistent with established KDD phases.

Results showed that the districts with the highest assault related incidents during the 2020-2024 period were: Independencia, La Victoria and Miraflores, showing persistent spatial patterns in consistency with Lima urban structure. These discoveries demonstrate the capacity of georeferenced text mining to identify critical zones from digital press information, consolidating a replicable methodology for urban crime analysis and the urban security decision-making support, not only in Lima, but in other cities in Peru and around the world.

Keywords: text mining, web scraping, georeferencing, spatial analysis, robberies, KDD.

Capítulo 1: Introducción

1.1 Antecedentes

Hablar de robos en el Perú, es un tema controversial que ha venido de años. El presente estudio, tratará de un análisis de datos entre los años 2020 a 2024 en Lima, con el fin de ver la contribución en la problemática social. Durante los años 2020 a 2021, los robos disminuyeron notablemente debido a las restricciones de movilidad por la pandemia. A partir de 2022, con el retorno progresivo a la actividad presencial, los robos aumentaron y se concentraron principalmente en robos personales en la vía pública. Los delitos que explican la mayor parte de la victimización son el robo de celulares, seguido del robo de dinero y carteras, generalmente sin uso de armas de fuego. En menor proporción aparecen los robos a viviendas y los robos de vehículos o autopartes, que son más focalizados. En 2024, aunque hay señales de ligera reducción en denuncias, el perfil delictivo sigue dominado por robos rápidos, oportunistas y de bajo monto, dirigidos a transeúntes y usuarios del transporte público. A continuación, se muestra cómo ha variado la tasa de víctimas de robo o intento de robo (por cada 100 habitantes de 15 años y más) en Lima Metropolitana:

Cuadro 1.1: Tasa de víctimas de robo o intento de robo (por cada 100 habitantes de 15 años y más) en Lima Metropolitana (enero 2020–diciembre 2024)

Año	Robo o intento de robo (total)	Variación vs. año anterior
2020	21.1	-
2021	14.9	-6.2
2022	18.9	4
2023	20.8	1.9
2024	19	-1.8

En relación a la situación actual mencionada, mediante la aplicación de herramientas de Big Data como web Scraping para la extracción de datos, se busca apoyar en la solución de dicha problemática.

En el caso del web scraping implica capturar y extraer datos de páginas web, generalmente utilizando herramientas automatizadas como bots o rastreadores web. Estos datos pueden ser utilizados para una amplia gama de aplicaciones, como la creación de modelos de lenguaje para chatbots o el análisis de patrones y tendencias en la información publicada en la web.

La minería de texto, por otro lado, es una técnica que se basa en el procesamiento del lenguaje natural (NLP), el cual es parte de la Inteligencia Artificial (IA) y de la Minería de Datos (DM) y se combinan con técnicas de Machine Learning enfocadas en la

clasificación o agrupamiento como lo son las redes neuronales. Esta técnica se utiliza para extraer información de alta calidad de grandes conjuntos de datos, como noticias, correos electrónicos o tweets. Algunas aplicaciones típicas de minería de texto incluyen análisis de sentimientos, extracción de información y modelado de temas. He aquí algunas diferencias entre la Minería de Texto y la Minería de Datos (ver Figura 1.1).



Figura 1.1: Diferencia entre minería de texto y minería de datos

Nota. Adaptado de Text mining: Analyzing unstructured data, por Project Guru (s. f.), <https://www.projectguru.in/text-mining-analyzing-unstructured-data/>

En resumen, el web scraping se utiliza para obtener datos de sitios web, mientras que la minería de texto se utiliza para procesar y analizar esos datos, extrayendo información de alta calidad y descubriendo patrones y tendencias en la información. A continuación, se presenta en la Figura Y unos gráficos que muestran el proceso de web scraping.

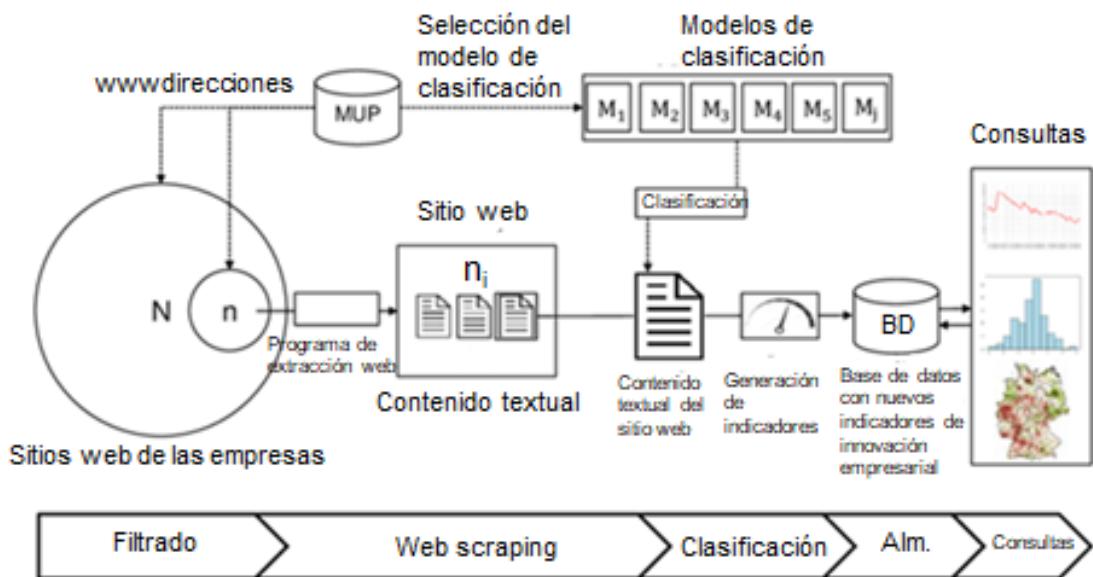


Figura 1.2: Flujo de trabajo: web scraping, clasificación de texto, generación de indicadores de innovación, almacenamiento y consulta

Nota. Adaptado de Exploring the use of text mining for innovation indicator development, por J. Kinne, 2018, Technology Innovation Management Review, 8(6), 43–49 (<https://doi.org/10.22215/timreview/1160>)

1.2 Planteamiento del problema

La problemática de los robos en Lima, a través de un análisis de datos apoyado con web scraping y minería de texto puede ayudar de manera concreta a prevenir, focalizar y reducir dichos robos. En términos prácticos, se contribuiría en:

1. Desconocimiento de lo reporte de noticias de robos en Lima en medios digitales. Actualmente, existe una gran cantidad de noticias sobre robos publicadas en medios digitales, portales web y plataformas informativas, las cuales se encuentran dispersas y no estructuradas. Esta fragmentación impide contar con una visión integral del fenómeno delictivo, dificultando la identificación de patrones espaciales, temporales y semánticos asociados a los robos en Lima. La ausencia de un sistema de análisis que consolide y procese estos textos limita el aprovechamiento de una fuente de información potencialmente valiosa para la comprensión del contexto urbano de la delincuencia.
2. Validez de la información, es decir en qué medida con símiles a las fuentes oficiales. Si bien las noticias web constituyen una fuente rica en descripciones y detalles contextuales, su validez frente a los registros oficiales de criminalidad no siempre está claramente establecida. Existen posibles discrepancias en cuanto a frecuencia, localización geográfica y tipología de los robos reportados, debido a sesgos editoriales, cobertura mediática selectiva o falta de

estandarización en la información. Analizar el grado de correspondencia entre los datos extraídos de medios digitales y las estadísticas oficiales permitirá evaluar la confiabilidad de estas fuentes alternativas y su potencial uso complementario en estudios de seguridad ciudadana.

3. Cómo las noticias web repercuten en la percepción de la población (en contraposición a los datos oficiales). Las noticias sobre robos difundidas en medios digitales influyen directamente en la percepción de inseguridad de la población, muchas veces amplificando el temor ciudadano más allá de lo reflejado por los datos oficiales. El lenguaje utilizado, la reiteración de ciertos eventos y la focalización en zonas específicas pueden generar una construcción mediática del delito que no siempre coincide con la realidad estadística. Comprender esta brecha entre percepción social y registros oficiales resulta fundamental para interpretar el impacto de la información digital en la opinión pública y en la toma de decisiones tanto individuales como institucionales.

1.3 Objetivos

1.3.1 General

Identificar patrones espacio temporales en noticias ocurridas en Lima Perú sobre robos durante los años 2020-2024.

1.3.2 Específicos

- Describir comportamientos de los lugares y temporalidades de los robos en Lima.
- Clasificar los comportamientos de los robos considerando el espacio y tiempo.
- Georreferenciar los registros de robos clasificados.

1.4 Justificación

La justificación del presente trabajo es poder contribuir en la problemática social de los robos en Lima, a través de las siguientes acciones:

1. Detección temprana de zonas y modalidades: El web scraping de noticias, redes sociales y foros vecinales permite identificar lugares, horarios y tipos de robo que aparecen con mayor frecuencia, incluso antes de que se reflejen en denuncias oficiales.
2. Análisis de patrones y evolución del delito: La minería de texto permite clasificar automáticamente los robos por modalidad (arrebato, asalto, moto lineal, transporte público) y analizar cómo cambian en el tiempo, ayudando a priorizar recursos policiales según tendencias reales.

3. Mejor asignación de recursos públicos: Con mapas de calor y análisis espacio-temporal, las autoridades pueden optimizar patrullajes, cámaras y operativos en zonas críticas, en lugar de aplicar estrategias generales poco efectivas.
4. Reducción de víctimas que no denuncian: Analizar texto en redes sociales y medios digitales permite capturar experiencias no reportadas, reduciendo el sesgo de las estadísticas basadas solo en denuncias.
5. Diseño de políticas públicas basadas en evidencia: Los resultados pueden usarse para campañas de prevención focalizadas (por ejemplo, horarios y zonas de mayor robo de celulares) y para evaluar qué medidas realmente funcionan.

En la actualidad, el análisis para datos textuales presenta problemas al momento de realizar análisis, pues existen escasas herramientas para su procesamiento y tratamiento de forma generalizada. Sin embargo, la importancia de los datos textuales en la actualidad radica en su capacidad para extraer información valiosa a partir del análisis de textos en lenguaje natural, lo que permite comprender patrones de comportamiento, preferencias, y opiniones. Estos datos son fundamentales en diversos ámbitos, como la investigación académica, el análisis del discurso, la verificación de información, la toma de decisiones en el ámbito empresarial, etc. El análisis de datos textuales posibilita la detección de vocabulario específico, el estudio de la argumentación, la evaluación institucional, y la comprensión de las preferencias, patrones de compra de los consumidores, etc. Además, en un contexto de "infodemia" (sobreabundancia de información), la verificación de la información se vuelve crucial, y el análisis de datos textuales contribuye a contrastar la información y aportar mayor objetividad a una investigación.

Los datos textuales incluyen información que se puede utilizar de diferentes formas para una exploración exhaustiva, ya sea desde recuentos de palabras hasta la georreferenciación mediante contextos, se refiere al proceso de asociar datos con una ubicación geográfica específica. Es así que, para este presente estudio o trabajo, la georreferenciación es una herramienta útil que se usará el cual estudia los fenómenos con relación espacial (cartográficamente y geográficamente) por métodos geoestadísticos (Giraldo, 2002), los cuales permiten identificar de donde en específico proviene la información. En este caso, los datos textuales para poder ser georreferenciados necesitan ser extraídos mediante un procesamiento donde se requiere de métodos estadísticos que coadyuven en su georreferenciación de los datos textuales. Es hasta ahí, la georreferenciación, será el alcance del presente estudio.

Asimismo, indicar que para lograr el objetivo de este estudio o trabajo se requiere de un proceso que ayude a obtener los datos (extracción) y de un tratamiento de la información previo para el análisis, el tratamiento de la información, es decir en síntesis un proceso KDD (Knowledge Discovery in Databases).

Capítulo 2: Estado del arte

El delito de robo se encuentra tipificado en el Código Penal peruano. De acuerdo con el Decreto Legislativo N.^º 635, se establece que “El que, para procurarse para sí o para otro un provecho ilícito, sustraе un bien mueble, total o parcialmente ajeno, empleando violencia contra la persona o amenaza de un peligro inminente para su vida o integridad física, será reprimido con pena privativa de libertad no menor de tres ni mayor de ocho años” (Perú, 1991, art. 188). Esta definición normativa delimita jurídicamente el fenómeno analizado en el presente estudio, proporcionando el marco legal para la clasificación de los hechos reportados en las noticias digitales.

Asimismo, según la Real Academia Española (s. f.), el término robo se define como la acción y efecto de robar, así como la cosa sustraída. En este sentido, el verbo robar alude a quitar o tomar para sí lo ajeno mediante violencia o fuerza, o incluso a hurtar de cualquier modo. Asimismo, puede referirse al hecho de despojar a una persona de aquello que legal o legítimamente posee, causando un perjuicio con ánimo de lucro. De manera más amplia, el término también puede emplearse para describir la obtención ilegítima de bienes o beneficios, e incluso para aludir metafóricamente a conseguir algo sin merecerlo o sin realizar el esfuerzo correspondiente (Real Academia Española, s. f.).

Por lo tanto, a partir de estas definiciones queda claro que el alcance de robos a ser analizados en este trabajo, son las que no conllevan a un crimen o asesinato de por medio.

A continuación, se presentan algunos conceptos aterrizados al presente trabajo el cual concluye en la aplicación de la Georreferenciación a la data textual usando como herramienta el Web Scraping para la recolección de la data.

2.1 Minería de datos

La minería de datos constituye una de las etapas centrales dentro del proceso de Descubrimiento de Conocimiento en Bases de Datos (Knowledge Discovery in Databases, KDD). Fayyad, Piatetsky-Shapiro y Smyth (1996) definen el KDD como un proceso no trivial orientado a identificar patrones en los datos que sean válidos, novedosos, potencialmente útiles y comprensibles desde una perspectiva humana. En este sentido, los autores enfatizan que el KDD trasciende la simple aplicación de algoritmos de minería de datos, ya que comprende una secuencia estructurada de etapas que incluyen la selección de datos relevantes, su limpieza y preprocesamiento, la transformación de los mismos, la aplicación de técnicas de minería de datos y, finalmente, la interpretación y evaluación de los resultados obtenidos. De esta manera, el objetivo fundamental del proceso KDD es transformar grandes volúmenes de datos en conocimiento significativo que contribuya a la toma de decisiones y a la comprensión de fenómenos complejos.

Este proceso de KDD aplicado a la presente investigación, se implementó de la siguiente manera:

1. Selección de datos: La fase de selección de datos constituye la etapa inicial del proceso de Descubrimiento de Conocimiento en Bases de Datos (KDD) y se orienta a identificar, integrar y seleccionar la información relevante proveniente de una o múltiples fuentes, en función del objetivo del análisis. Esta etapa permite delimitar el universo de datos, reducir la complejidad del conjunto inicial y determinar el subconjunto que será utilizado dentro del proceso de descubrimiento de conocimiento. En este sentido, Fayyad et al. (1996) explican que la selección de datos implica decidir qué porción del conjunto total será empleada en el proceso de descubrimiento de conocimiento.

2. Preprocesamiento: La fase de preprocesamiento de datos tiene como finalidad corregir errores, inconsistencias y ruido presentes en la información. Esta etapa incluye el manejo de valores faltantes, la eliminación de registros duplicados y la corrección de errores estructurales. En el contexto de la minería de textos, el preprocesamiento implica, además, la eliminación de contenido irrelevante como spam y la normalización lingüística. En este sentido, Han et al. (2012) señalan que la calidad de los datos constituye un factor crítico en el proceso de minería, siendo necesario aplicar técnicas de preprocesamiento para eliminar ruido y valores inconsistentes.

3. Transformación de datos: La fase de transformación de datos consiste en convertir la información en un formato adecuado para la aplicación de algoritmos de minería. Esta etapa puede incluir procesos como la normalización, agregación, reducción de dimensionalidad e ingeniería de variables. En el ámbito de la minería de textos, la transformación comprende tareas como la tokenización, lematización y representación vectorial mediante técnicas como TF-IDF o embeddings. En este sentido, Kantardzic (2011) explica que la transformación de datos tiene como propósito mejorar la eficiencia del proceso analítico y la calidad de los patrones descubiertos.

4. Minería de datos: La fase de minería de datos constituye el núcleo del proceso KDD, ya que en ella se aplican algoritmos con el propósito de identificar patrones significativos en los datos. Entre las técnicas más empleadas se encuentran la clasificación, el análisis de conglomerados (clustering), las reglas de asociación y la detección de anomalías. En el ámbito de la minería de textos, esta etapa puede incluir métodos como el modelado de temas, la clasificación automática de documentos y el análisis de sentimiento. En este sentido, Fayyad et al. (1996) explican que la minería de datos corresponde a la etapa del proceso KDD en la cual se aplican algoritmos para extraer patrones interesantes a partir de los datos.

5. Evaluación e interpretación: La fase de evaluación e interpretación tiene como propósito validar la utilidad y el significado de los patrones descubiertos durante la etapa de minería de datos. Esta etapa comprende la validación estadística de los resultados, el análisis de su interpretabilidad y la determinación de su relevancia dentro del dominio

de estudio. En este sentido, Fayyad et al. (1996) señalan que no todos los patrones identificados resultan necesariamente útiles, por lo que el proceso de evaluación permite determinar cuáles poseen verdadero interés y valor para la toma de decisiones.

6. Presentación y uso del conocimiento: La fase de presentación del conocimiento tiene como propósito comunicar los resultados obtenidos de manera clara y comprensible, con el fin de apoyar la toma de decisiones. Para ello, se emplean herramientas como visualizaciones, reportes analíticos, paneles interactivos (dashboards), mapas y otros recursos gráficos que facilitan la interpretación de los hallazgos. En este sentido, Han et al. (2012) señalan que el conocimiento descubierto debe representarse de forma comprensible para el usuario final, asegurando que los patrones identificados puedan ser interpretados y utilizados adecuadamente dentro del contexto del dominio de aplicación.

El desarrollo de la minería de datos como disciplina científica se encuentra estrechamente vinculado al concepto de Descubrimiento de Conocimiento en Bases de Datos (Knowledge Discovery in Databases, KDD). En este contexto, Fayyad et al. (1996) proponen una de las definiciones más influyentes y ampliamente citadas en la literatura, describiendo el KDD como un proceso no trivial orientado a identificar patrones en los datos que sean válidos, novedosos, potencialmente útiles y comprensibles. Esta definición enfatiza que el descubrimiento de conocimiento no se limita a la aplicación de algoritmos, sino que constituye un proceso integral que abarca desde la selección de datos relevantes hasta la interpretación final de los patrones identificados.

Dentro de este marco conceptual, los autores señalan que la minería de datos representa la fase central del proceso KDD, ya que es en esta etapa donde se aplican métodos computacionales y estadísticos para extraer regularidades significativas a partir de grandes volúmenes de información. Mientras que el KDD comprende el flujo metodológico completo para transformar datos en conocimiento, la minería de datos constituye el núcleo analítico encargado de generar patrones, modelos y estructuras interpretables, Fayyad et al. (1996).

En cuanto a las tareas fundamentales de la minería de datos, Fayyad et al. (1996) explican que estas pueden adoptar diversas formas según el objetivo del análisis. Una de las principales es la clasificación, la cual consiste en construir modelos predictivos capaces de asignar categorías discretas a nuevas observaciones basándose en ejemplos previamente etiquetados. Este enfoque resulta especialmente relevante cuando se requiere diferenciar entre tipos de eventos o fenómenos, permitiendo automatizar procesos de identificación y categorización.

Otra tarea destacada es la regresión, orientada a modelar relaciones funcionales entre variables y predecir valores numéricos continuos. A diferencia de la clasificación, que trabaja con clases discretas, la regresión se centra en estimaciones cuantitativas, siendo particularmente útil cuando se busca anticipar magnitudes o intensidades asociadas a determinados fenómenos, Fayyad et al. (1996).



Asimismo, los autores describen el agrupamiento o clustering como una función esencial de la minería de datos. Esta tarea tiene como propósito identificar estructuras internas en los datos mediante la formación de grupos homogéneos sin necesidad de categorías predefinidas. El clustering resulta especialmente pertinente en estudios exploratorios, donde el objetivo no es confirmar hipótesis previas, sino descubrir patrones emergentes o segmentaciones latentes, Fayyad et al. (1996).

De igual manera, Fayyad et al. (1996) incluyen la caracterización o resumificación como otra tarea relevante, cuyo objetivo es generar representaciones compactas y comprensibles de conjuntos de datos complejos. Esta función puede materializarse mediante descripciones agregadas, extracción de atributos significativos o técnicas de visualización que faciliten la interpretación humana de los resultados obtenidos.

En conjunto, estas tareas evidencian que la minería de datos no constituye un procedimiento único, sino un conjunto diverso de enfoques analíticos orientados a distintos propósitos: predecir, describir, segmentar y sintetizar información. Su aporte resulta fundamental en investigaciones que buscan transformar datos extensos, heterogéneos y complejos en patrones interpretables que respalden la toma de decisiones y la comprensión de fenómenos analizados, Fayyad et al. (1996).

Según Villa Monte (2023), en el material correspondiente al Tema 2 de la asignatura Minería de Datos de la Universidad Internacional de Valencia, se establecen definiciones conceptuales que sirven como base para el desarrollo del presente estudio.

- Aprendizaje supervisado establece una correspondencia entre las entradas y las salidas del sistema, donde la base de conocimientos del sistema está formada por ejemplos etiquetados a priori.
- Aprendizaje no supervisado el proceso de modelado se lleva a cabo sobre un conjunto de ejemplos formados únicamente por entradas al sistema, sin conocer su clasificación correcta.
- Aprendizaje por refuerzo el algoritmo aprende observando el mundo que le rodea y con un continuo flujo de información en las dos direcciones realizando un proceso de ensayo error, y reforzando aquellas acciones que reciben una respuesta positiva en el mundo.

Por otro lado, la minería de datos y la minería de texto son disciplinas relacionadas que comparten objetivos y técnicas, pero se enfocan en tipos de datos diferentes; en este sentido es necesario primero tener claro las definiciones de ambos.

La minería de datos es definida como la fase del proceso KDD en la cual se aplican algoritmos específicos para extraer patrones a partir de los datos. En este sentido, Fayyad et al. (1996) señalan que su función principal consiste en descubrir regularidades potencialmente útiles dentro de grandes volúmenes de información. Asimismo, los autores enfatizan que la minería de datos no constituye un proceso completo por sí mismo, sino que representa una etapa esencial dentro del flujo

metodológico del descubrimiento de conocimiento, la cual debe complementarse con fases posteriores de interpretación y evaluación para otorgar significado a los patrones identificados.

Diversos autores definen la minería de textos como el proceso de extracción de conocimiento a partir de datos textuales no estructurados. Hearst (1999) enfatiza que este proceso implica técnicas específicas de procesamiento del lenguaje natural, diferenciándose de la minería de datos tradicional aplicada a datos estructurados.

Sobre la base de la definición de minería de datos y minería de texto, y el proceso de KDD, se ha elaborado el siguiente cuadro comparativo:

Cuadro 2.1: Comparativo entre Minería de datos y Minería de textos

Aspecto / Etapa del proceso KDD	Minería de Datos (Data Mining)	Minería de Textos (Text Mining)	Similitudes / Diferencias clave
Tipo de dato principal	Datos estructurados y semiestructurados (tablas, bases de datos, registros numéricos).	Datos no estructurados (documentos, noticias web, textos libres, redes sociales).	Diferencia central: el texto requiere estructuración previa mediante NLP.
Selección de datos	Se seleccionan variables relevantes desde bases de datos organizadas.	Se seleccionan documentos textuales relevantes (corpus), generalmente de fuentes web o repositorios digitales.	Similaridad: ambos requieren definir fuentes y criterios de inclusión.
Preprocesamiento y limpieza	Limpieza de valores faltantes, duplicados, errores, normalización numérica.	Limpieza lingüística: eliminación de stopwords, corrección, tokenización, lematización.	Diferencia: en textos se requiere procesamiento del lenguaje natural.
Transformación de datos	Conversión a formatos analizables (escalamiento, reducción dimensional, codificación).	Conversión del texto en representaciones numéricas: TF-IDF, embeddings, bolsas de palabras.	Similaridad: ambos transforman datos a estructuras utilizables por algoritmos.
Minería de datos (etapa central)	Aplicación de algoritmos para extraer patrones (clasificación, clustering, regresión).	Aplicación de algoritmos similares, pero sobre variables textuales extraídas o semánticas.	Similaridad: comparten tareas analíticas; diferencia: el input textual es más complejo.

Interpretación y evaluación	Evaluación estadística de modelos, validación de patrones, métricas numéricas.	Evaluación adicional de coherencia semántica y relevancia contextual del lenguaje.	Diferencia: el significado del texto requiere interpretación lingüística.
Conocimiento útil (output)	Modelos predictivos, reglas de asociación, segmentaciones, patrones cuantitativos.	Descubrimiento de temas, tendencias, relaciones semánticas, eventos descritos en textos.	Similaridad: ambos generan conocimiento; diferencia: text mining revela contenido temático.
Herramientas comunes	Estadística, machine learning, bases de datos, modelos supervisados/no supervisados.	NLP, machine learning, análisis semántico, topic modeling, extracción de entidades.	Text mining incorpora herramientas lingüísticas adicionales.
Aplicaciones típicas	Finanzas, industria, fraude, salud, logística, predicción cuantitativa.	Ánalysis de noticias, opinión pública, redes sociales, inteligencia criminal basada en texto.	Diferencia: text mining se orienta más a fuentes discursivas.
Relación con KDD	Data mining es la aplicación de algoritmos para extraer patrones desde datos estructurados.	Text mining aplica minería sobre datos textuales previamente estructurados mediante NLP.	Ambos son parte del proceso KDD, pero con distinta naturaleza de datos.

Asimismo, es importante mencionar para este estudio las aplicaciones integradas entre ambas minerías. A menudo, la minería de datos y la minería de texto se combinan para resolver problemas complejos que involucran ambos tipos de datos. Por ejemplo:

- **Análisis de Opiniones en Redes Sociales:** El análisis de opiniones en redes sociales, también denominado análisis de sentimiento, constituye un campo de estudio orientado a la identificación y análisis computacional de opiniones, emociones y actitudes expresadas en texto. En este sentido, Liu (2012) lo define como el estudio computacional de las opiniones y sentimientos presentes en datos textuales, destacando su utilidad para determinar la orientación subjetiva hacia entidades o temas específicos y apoyar procesos de toma de decisiones basados en grandes volúmenes de información generada por los usuarios.
- **Sistemas de Recomendación:** Los sistemas de recomendación se definen como herramientas y técnicas de software diseñadas para proporcionar sugerencias personalizadas a los usuarios sobre productos, servicios o contenidos que pueden resultar de su interés. En este sentido, Ricci et al. (2011) explican que estos sistemas funcionan mediante el análisis de preferencias,



comportamientos previos e interacciones del usuario, con el propósito de reducir la sobrecarga de información y facilitar la toma de decisiones en entornos digitales caracterizados por una amplia oferta de opciones.

La convergencia de la minería de datos y minería de texto es evidente en el análisis de Big Data, donde se integran datos estructurados y no estructurados para proporcionar una visión más completa y detallada. La minería de datos puede beneficiarse de la minería de texto para enriquecer los conjuntos de datos estructurados con información derivada de textos, mejorando así la precisión y la profundidad del análisis. En resumen, aunque la minería de datos y la minería de texto se enfocan en diferentes tipos de datos y utilizan técnicas específicas para cada uno, ambas están estrechamente relacionadas y a menudo se complementan para ofrecer análisis más completos y detallados.

2.2 Minería de texto

Según lo indicado en el acápite anterior, la Minería de texto busca extraer información y patrones de un gran volumen de datos textuales, y para ello se basa en métodos de la minería de datos, aprendizaje automático, estadística, y procesamiento de lenguaje natural (PLN) para explorar. El objetivo principal es transformar el texto en datos estructurados para análisis, descubrimiento de patrones, extracción de entidades, categorización, y otras tareas que permitan la toma de decisiones basada en datos. Por lo tanto, son funciones de la Minería de texto:

La minería de textos surge como una extensión de la minería de datos orientada específicamente al análisis de información no estructurada expresada en lenguaje natural. Hearst (1999) sostiene que, a diferencia de los datos estructurados tradicionales, los documentos textuales requieren procesos adicionales de representación y técnicas propias del procesamiento del lenguaje natural para posibilitar la extracción de conocimiento. En este sentido, la autora enfatiza que la minería de textos involucra tareas destinadas a identificar patrones y estructuras significativas dentro de grandes colecciones documentales.

Hearst (1999) señala que la minería de textos integra diversas tareas provenientes del procesamiento del lenguaje natural y la recuperación de información, entre ellas la extracción de información, que permite identificar entidades y relaciones dentro de los documentos y transformar texto no estructurado en representaciones analizables.

Otra tarea asociada a la minería de textos es la identificación de temas predominantes en colecciones documentales amplias. Hearst (1999) señala que, mediante técnicas como el agrupamiento y otros métodos de análisis textual, es posible descubrir la estructura temática presente en grandes volúmenes de información. Este enfoque facilita la identificación de patrones y relaciones relevantes dentro de medios digitales.

Hearst (1999) señala que la clasificación de documentos constituye una de las tareas frecuentes en el ámbito de la minería de textos. Esta tarea consiste en asignar categorías o etiquetas a los documentos en función de su contenido, lo que permite organizar y filtrar información de manera automatizada. En contextos aplicados, la clasificación resulta especialmente útil para diferenciar documentos vinculados a fenómenos específicos, como distintos tipos de delitos reportados en noticias digitales.

Hearst (1999) incluye entre las tareas asociadas a la minería de textos la agrupación o clustering de documentos, la cual permite organizar textos en grupos homogéneos sin recurrir a categorías predefinidas. Esta técnica resulta especialmente útil en análisis exploratorios, ya que facilita la identificación de estructuras latentes dentro del corpus textual y el reconocimiento de subconjuntos temáticos relevantes.

Así también, Hearst (1999) señala la estrecha relación entre la minería de textos y la recuperación de información, destacando que el análisis textual puede complementar los sistemas tradicionales de búsqueda. En este contexto, la autora explica que las técnicas de procesamiento del lenguaje natural permiten ir más allá de la simple coincidencia de palabras clave, facilitando una mejor identificación de documentos relevantes dentro de grandes colecciones textuales.

Asimismo, Hearst (1999) hace referencia a tareas relacionadas con el análisis textual, entre ellas el resumen automático, cuyo objetivo es generar representaciones condensadas del contenido de uno o varios documentos. Esta técnica facilita la interpretación humana cuando se trabaja con grandes colecciones de información y contribuye a mitigar la sobrecarga informativa en entornos digitales.

Finalmente, Hearst (1999) plantea que la minería de textos se orienta al análisis estructural de grandes colecciones documentales, integrando técnicas provenientes del procesamiento del lenguaje natural y la recuperación de información. En este marco, las tareas asociadas a la minería de textos permiten transformar documentos en representaciones analizables, identificar patrones temáticos, clasificar y agrupar textos, extraer información relevante y generar estructuras que faciliten su interpretación. De este modo, la minería de textos se consolida como una herramienta fundamental para el análisis de información textual en diversos contextos aplicados.

2.3 Web scraping

El web scraping, también denominado extracción automatizada de datos web, constituye una técnica empleada para la recopilación de información disponible en internet, especialmente cuando los datos no se presentan en formatos estructurados descargables. Mitchell (2018) explica que el web scraping consiste en el uso de programas informáticos que navegan por páginas web y extraen automáticamente información de su contenido, transformándola en conjuntos de datos reutilizables para análisis posterior. Según la autora, esta técnica permite recuperar información de fuentes digitales heterogéneas, como noticias, catálogos o plataformas sociales,



facilitando la construcción de bases de datos a partir de información originalmente diseñada para ser consumida por humanos.

En este sentido, el web scraping se ha convertido en una herramienta clave dentro de investigaciones basadas en grandes volúmenes de datos, ya que permite acceder a fuentes actualizadas en tiempo real y recopilar información de manera sistemática. Su uso resulta particularmente relevante en contextos donde se requiere analizar dinámicas sociales o fenómenos urbanos reflejados en medios digitales, como ocurre con noticias georreferenciadas o reportes de delitos en portales web.

Clasificación del Web Scraping según el tipo de extracción:

Mitchell (2018) señala que el web scraping puede implementarse mediante distintos enfoques técnicos, dependiendo de la estructura del sitio web y del tipo de información que se desea extraer. A partir de estas consideraciones, es posible distinguir diversas modalidades operativas según el nivel de complejidad y la interacción requerida con la página.

En primer lugar, se encuentra el scraping basado en HTML estático, en el cual el contenido se extrae directamente del código fuente de páginas web tradicionales. Este tipo de scraping es común en sitios donde la información se encuentra disponible de forma inmediata en el documento HTML y puede ser recuperada mediante selectores o patrones estructurales.

En segundo lugar, se identifica el scraping de contenido dinámico, característico de páginas modernas que cargan información mediante JavaScript o solicitudes asíncronas. En estos casos, es necesario emplear herramientas capaces de simular la interacción humana o reproducir el comportamiento de un navegador, ya que los datos no aparecen directamente en el código inicial de la página.

Finalmente, Mitchell (2018) describe también escenarios de scraping que implican la interacción con estructuras específicas de los sitios web, tales como formularios, motores de búsqueda internos o páginas con múltiples niveles de navegación. En estos casos, el proceso de extracción requiere recorridos automatizados más complejos para recolectar información de manera exhaustiva.

Esta clasificación permite comprender que el web scraping no constituye una técnica única, sino un conjunto de estrategias adaptadas a diferentes arquitecturas web y necesidades de investigación.

Relación entre Web Scraping y Minería de Textos:

El web scraping mantiene una relación directa con la minería de textos, particularmente en investigaciones donde los datos de interés se presentan en forma discursiva o narrativa. En este marco, el scraping representa la etapa inicial de adquisición de datos

textuales desde internet, mientras que la minería de textos constituye la fase analítica encargada de transformar esos documentos no estructurados en conocimiento útil.

Hearst (1999) sostiene que la minería de textos se orienta al análisis estructural de grandes colecciones documentales con el fin de identificar patrones, relaciones y estructuras relevantes dentro de información no estructurada. No obstante, para que dicho análisis sea viable en entornos digitales contemporáneos, resulta indispensable contar previamente con un corpus textual organizado y sistemáticamente recopilado. En este contexto, las técnicas de web scraping descritas por Mitchell (2018) permiten automatizar la extracción de contenidos desde páginas web, facilitando la construcción de bases de datos textuales a partir de información originalmente diseñada para el consumo humano. De este modo, el web scraping se configura como una etapa preliminar necesaria que posibilita la aplicación posterior de tareas propias de la minería de textos, tales como clasificación de documentos, detección de temas, extracción de entidades o análisis de tendencias.

En investigaciones aplicadas, como el análisis de noticias sobre robos en Lima, el proceso suele estructurarse de manera secuencial: en una primera etapa se recopilan las noticias mediante técnicas de web scraping (Mitchell, 2018); posteriormente se realiza el procesamiento lingüístico del corpus textual, lo que permite aplicar tareas propias de la minería de textos orientadas a la identificación de patrones y relaciones relevantes (Hearst, 1999). Finalmente, los resultados pueden integrarse con herramientas de georreferenciación para analizar la dimensión espacial del fenómeno estudiado.

2.4 Georreferenciación

La georreferenciación constituye un concepto fundamental dentro de los Sistemas de Información Geográfica (SIG) y el análisis espacial, ya que permite asociar información de diversa naturaleza con una ubicación específica en el espacio geográfico. Hill (2006) explica que este proceso implica establecer vínculos entre datos o recursos y su localización geográfica, posibilitando su organización, visualización y análisis en función del espacio. Este enfoque resulta especialmente relevante cuando se trabaja con datos originalmente no espaciales, como documentos textuales o registros administrativos, que requieren ser vinculados a coordenadas o unidades territoriales para su aprovechamiento analítico.

De manera complementaria, Hackeloeer et al. (2014) describen la georreferenciación como un término amplio que engloba diversos métodos orientados a la identificación única de objetos o eventos geográficos. Los autores señalan que georreferenciar no se limita a la asignación de coordenadas, sino que implica la aplicación de técnicas que permiten localizar entidades de forma inequívoca dentro de un marco espacial definido. Esta perspectiva amplía el concepto hacia un enfoque metodológico general aplicable tanto a datos estructurados como no estructurados.



Clasificación de la referencia espacial:

Desde un marco normativo, la Organización Internacional de Normalización distingue dos enfoques principales para la referencia espacial: la referenciación basada en coordenadas y la referenciación basada en identificadores geográficos. La norma ISO 19111:2007 establece los principios para la referenciación espacial mediante sistemas de coordenadas, mientras que la norma ISO 19112:2003 regula la referenciación por identificadores geográficos, tales como nombres de lugares o unidades administrativas. Esta distinción permite comprender que la georreferenciación puede realizarse de forma directa, cuando se dispone de coordenadas explícitas, o de forma indirecta, cuando es necesario transformar descriptores espaciales en ubicaciones precisas.

Relación entre georreferenciación y minería de textos:

En investigaciones que emplean fuentes textuales digitales, como noticias web o redes sociales, la georreferenciación se vincula con la minería de textos mediante técnicas de reconocimiento de ubicaciones. En este contexto, Leidner (2008) explica que el geoparsing consiste en identificar referencias geográficas en el lenguaje natural, especialmente nombres de lugares o topónimos, y tratarlas como entidades espaciales que deben ser posteriormente localizadas de forma precisa. Este proceso comprende tanto el reconocimiento automático de menciones geográficas como su posterior resolución o desambiguación espacial, con el objetivo de asignar una representación geoespacial adecuada.

En estudios recientes, Tao et al. (2022) señalan que el reconocimiento de entidades geográficas en textos implica identificar expresiones lingüísticas asociadas a ubicaciones y tratarlas como entidades espaciales que puedan ser posteriormente localizadas y analizadas espacialmente. Estos autores destacan que este proceso, conocido como reconocimiento de entidades geográficas (geographic named entity recognition), permite vincular grandes volúmenes de datos textuales con representaciones geoespaciales útiles para su integración en sistemas de información geográfica y metodologías de minería de datos espaciales.

En consecuencia, la relación entre georreferenciación y minería de textos se materializa a través del geoparsing, proceso mediante el cual las referencias espaciales presentes en textos no estructurados son identificadas y vinculadas a representaciones geográficas formales (Leidner, 2008; Tao et al., 2022). En investigaciones aplicadas, como el análisis de noticias georreferenciadas sobre robos, esta integración permite transformar menciones a distritos, avenidas o zonas urbanas en elementos cartográficos susceptibles de análisis espacial. De este modo, la minería de textos aporta la extracción semántica de ubicaciones, mientras que la georreferenciación posibilita su representación espacial, facilitando la identificación de patrones territoriales en contextos urbanos y criminológicos.

2.5 Técnicas estadísticas para datos textuales

2.5.1 Tabla de frecuencias

Una tabla de frecuencias constituye una herramienta básica de la estadística descriptiva empleada para organizar y resumir datos. Freedman et al. (2007) explican que una tabla de frecuencias es un arreglo sistemático que presenta la distribución de un conjunto de observaciones mediante el conteo del número de veces que cada valor o categoría aparece en los datos. Este tipo de representación permite mostrar frecuencias absolutas o relativas, facilitando la identificación de patrones generales y la descripción preliminar del fenómeno analizado.

2.5.2 Tabla de contingencias

La tabla de contingencia constituye el insumo fundamental del análisis de correspondencias. Según Greenacre (2017), esta tabla está formada por frecuencias cruzadas que resumen la distribución conjunta de dos variables categóricas, permitiendo examinar la relación entre sus categorías. A partir de dicha matriz de datos, el análisis de correspondencias descompone la estructura asociativa mediante técnicas de reducción dimensional, generando representaciones gráficas que facilitan la interpretación de los patrones existentes.

2.5.3 Análisis de correspondencia

El análisis de correspondencias es una técnica estadística multivariante orientada a explorar y representar gráficamente las relaciones entre categorías de variables cualitativas a partir de tablas de contingencia. Greenacre (2017) explica que este método analiza las asociaciones entre filas y columnas mediante la descomposición de la estructura de dependencia, proyectando la información en un espacio geométrico de baja dimensión en el que las distancias entre puntos reflejan la similitud estadística entre categorías. De este modo, el análisis de correspondencias facilita la interpretación de estructuras subyacentes en datos categóricos, proporcionando representaciones gráficas que permiten identificar patrones y agrupamientos entre las modalidades observadas.

2.5.4 Análisis cluster

El análisis de clúster, también denominado análisis de conglomerados, es una técnica estadística multivariante cuyo objetivo es clasificar un conjunto de observaciones en grupos homogéneos, de modo que los elementos dentro de cada grupo presenten mayor similitud entre sí que respecto a los de otros grupos. Everitt et al. (2011) explican que esta metodología permite identificar estructuras subyacentes en los datos sin recurrir a categorías predefinidas, siendo especialmente útil en estudios exploratorios orientados al descubrimiento de patrones naturales o segmentaciones latentes. En este



sentido, el análisis de clúster constituye una herramienta relevante para la reducción de complejidad y la interpretación de relaciones internas en datos multivariantes.

En el presente trabajo se emplea el dendrograma como herramienta de clasificación jerárquica. El dendrograma es una representación gráfica en forma de árbol utilizada en el análisis de clúster jerárquico para ilustrar el proceso progresivo de agrupamiento de las observaciones. Everitt et al. (2011) señalan que esta representación muestra cómo los elementos individuales se fusionan en conglomerados sucesivos en función de su similitud, permitiendo visualizar la estructura jerárquica de los grupos y seleccionar el nivel de agrupación adecuado mediante el corte del árbol a distintas alturas. De este modo, el dendrograma constituye una herramienta interpretativa clave para comprender la formación y relación entre clústeres en un conjunto de datos multivariantes.

2.6 Otras herramientas para analizar datos textuales

Adicionalmente, es importante mencionar que existen otras herramientas con funciones parecidas o complementarias, las cuales se listan a continuación:

Procesamiento del Lenguaje Natural (NLP)

El Procesamiento del Lenguaje Natural (NLP) constituye una disciplina fundamental para el análisis automatizado de textos. Jurafsky y Martin (2023) describen el NLP como un campo interdisciplinario que integra lingüística computacional, inteligencia artificial y aprendizaje automático con el propósito de permitir que las computadoras procesen, interpreten y generen lenguaje humano. En este marco, el NLP proporciona las bases técnicas necesarias para tareas como la tokenización, lematización, análisis sintáctico y extracción semántica, consolidándose como un componente esencial en proyectos de minería de textos aplicada.

Extracción de Información

La extracción de información constituye una técnica fundamental para estructurar datos provenientes de textos no estructurados. Sarawagi (2008) describe la extracción de información como el proceso de identificar automáticamente entidades, relaciones y eventos específicos dentro de documentos textuales, transformando el contenido en datos organizados susceptibles de análisis posterior. Esta metodología resulta particularmente útil en el contexto de noticias digitales, donde es posible extraer ubicaciones, tipos de delito o actores involucrados.

Modelado de Temas

El modelado de temas constituye una técnica estadística orientada a identificar estructuras temáticas latentes en grandes colecciones de documentos. Blei (2012) explica que los modelos probabilísticos de temas permiten descubrir automáticamente conjuntos de palabras que tienden a aparecer conjuntamente en los textos, interpretándose como tópicos subyacentes dentro del corpus. Este enfoque facilita el análisis exploratorio y la identificación de patrones discursivos en documentos extensos.



Representación Vectorial del Texto (TF-IDF y Espacio Vectorial)

Una herramienta fundamental en la minería de textos es la representación numérica de documentos mediante el modelo de espacio vectorial. Salton y Buckley (1988) explican que este modelo permite representar los textos como vectores de términos ponderados, en los cuales esquemas como TF-IDF asignan mayor peso a aquellas palabras que resultan más discriminantes dentro del corpus. Esta representación facilita la aplicación de algoritmos de clasificación, agrupamiento (clustering) y recuperación de información.

Reconocimiento de Entidades Nombradas

El reconocimiento de entidades nombradas es una técnica orientada a la identificación y clasificación automática de entidades como personas, lugares, organizaciones u otras categorías semánticas dentro de un texto. Nadeau y Sekine (2007) explican que esta herramienta constituye un componente fundamental para estructurar información textual, facilitando tareas posteriores como la georreferenciación de noticias o el análisis de actores relevantes en eventos específicos.

Análisis de Sentimiento

El análisis de sentimiento constituye una herramienta ampliamente utilizada para examinar percepciones expresadas en textos digitales. Liu (2012) describe el análisis de sentimiento como el estudio computacional de opiniones, actitudes y emociones presentes en el lenguaje natural, especialmente en contenidos generados por usuarios en plataformas digitales. Este enfoque resulta útil para analizar la percepción ciudadana sobre fenómenos como la inseguridad o los robos en contextos urbanos.

2.7 Marco de referencia

Como antecedente relevante se encuentra la tesis titulada Métricas multivariantes textuales georreferenciadas: Caso aplicado al Estado de Veracruz 2020, elaborada por Hernández Salazar (2020). En dicho estudio se propone una metodología integral que comprende la extracción, el tratamiento estadístico y la georreferenciación de datos textuales provenientes de noticias web. La investigación demuestra la viabilidad de utilizar información no estructurada como insumo analítico, ofreciendo una alternativa metodológica para investigadores y analistas interesados en aprovechar fuentes textuales digitales en estudios espaciales.

En el ámbito de los Sistemas de Información Geográfica aplicados a la delincuencia, Espinoza-Ramírez, Nakano, Sánchez-Pérez y Arista-Jalife (2018) desarrollan un sistema computacional denominado GeoSecurityMX, el cual integra dispositivos móviles y geolocalización para la generación de indicadores delictivos. La metodología propuesta contempla diversas fases que incluyen la recopilación, procesamiento y visualización cartográfica de información, evidenciando la utilidad de la georreferenciación para la detección temprana de zonas de riesgo y el apoyo a la toma de decisiones en seguridad pública.



Asimismo, Moreno Jiménez y Grijalva Eternod (2022) proponen indicadores alternativos para la medición de la delincuencia urbana basados en la presencia efectiva de población, utilizando datos georreferenciados de Twitter y del Directorio Estadístico Nacional de Unidades Económicas (DENUE). Los resultados muestran que la incorporación de datos de actividad urbana permite identificar con mayor precisión poblaciones en riesgo, destacando el valor del uso de fuentes digitales como proxy de dinámicas sociales.

Desde la perspectiva del análisis espacial del delito, Harries (2001) establece las bases conceptuales del crime mapping, demostrando cómo la georreferenciación de eventos delictivos facilita la identificación de patrones espaciales, conglomerados y zonas calientes (hot spots). Complementariamente, Chainey, Tompson y Uhlig (2008) analizan la Estimación de Densidad Kernel (KDE) como técnica para la identificación de concentraciones delictivas, mostrando cómo la transformación de eventos puntuales en superficies continuas mejora la interpretación espacial del riesgo.

En cuanto a la minería de textos aplicada al crimen, Chen, Chung y Xu (2004) desarrollan un marco general para el uso de técnicas de procesamiento del lenguaje natural en el análisis criminal, evidenciando la utilidad de transformar textos no estructurados en variables cuantificables para apoyar la toma de decisiones. De manera complementaria, Wang, Gerber y Brown (2012) demuestran que los datos de redes sociales georreferenciados pueden emplearse para la predicción del crimen urbano, validando empíricamente el uso de textos digitales como sensores sociales.

Finalmente, Liu, Andris y Ratti (2010) integran información semántica con coordenadas geográficas en el análisis urbano, sentando bases metodológicas para la combinación de minería de textos y análisis espacial. Estos antecedentes refuerzan la pertinencia de integrar extracción textual, análisis estadístico multivariante y georreferenciación en el estudio de fenómenos delictivos en contextos urbanos.



Capítulo 3: Desarrollo del proyecto

3.1. Metodología

El presente proyecto se enmarca dentro del proceso de KDD, una metodología ampliamente utilizada para transformar grandes volúmenes de datos en información significativa y conocimiento útil. La adopción de esta metodología resulta especialmente pertinente debido a la naturaleza heterogénea y masiva de los datos periodísticos recolectados.

En este contexto, el proceso KDD constituye un marco estructurado para guiar las fases del trabajo, desde la recolección inicial de las noticias hasta la presentación de resultados. Su aplicación no solo permite organizar el trabajo metodológico, sino también garantizar la trazabilidad de las decisiones técnicas tomadas en cada etapa del desarrollo.

El KDD, en términos generales, se compone de diversas fases encadenadas: selección de datos, preprocesamiento y transformación, minería de datos, evaluación y presentación del conocimiento. Cada una de estas fases aporta valor en el camino hacia la construcción de conocimiento, asegurando que la información sea depurada, normalizada y analizada bajo criterios consistentes.

En el marco del presente proyecto, esta metodología se adaptó para el tratamiento de noticias periodísticas sobre los robos en Lima, con el objetivo identificar patrones mediante un análisis de datos (noticias) entre los años 2020 a 2024. Así, cada fase del KDD fue contextualizada a la naturaleza textual de los datos y al propósito central.

En este contexto, las fases del KDD se materializan a lo largo del desarrollo del trabajo, donde se detallan los procedimientos implementados en cada fase. De este modo, el esquema metodológico actúa como marco de referencia que articula las herramientas técnicas utilizadas, con los objetivos analíticos planteados, garantizando un flujo de procesamiento consistente y orientado al descubrimiento de conocimiento sobre los robos en Lima.

La implementación por cada una de las fases del KDD, aplicada al contexto de este trabajo de investigación, es la siguiente:

1. Selección de datos:

La fase de recopilación y almacenamiento de los datos se constituyó como un componente esencial del presente trabajo, al permitir la conformación de una base de noticias periodísticas coherente y estructurada, que funcionó como insumo para la aplicación posterior de las técnicas de preprocesamiento, minería de texto y georreferenciación. En este contexto, el objetivo no se centró exclusivamente en la



obtención de grandes volúmenes de información, sino en garantizar que los datos recolectados fueran pertinentes, trazables y adecuados para los procesos de análisis desarrollados en las etapas posteriores.

Es por tanto el objetivo de esta fase el recopilar las noticias en Lima entre los años 2020 y 2024, en este caso se seleccionó solo para los diarios El Comercio, Gestión, Correo y Ojo. Y para ello, se hizo uso de la metodología Web Scraping. El proceso de Web Scraping se realizó mediante la extracción de data de las noticias de las páginas web de los diarios mencionados, estructurándolos para obtener principalmente el título, el resumen y la fecha, y usando lenguaje de programación R. El código en RStudio del proceso Web Scraping fueron almacenados en la plataforma GitHub, ver Referencia I.

Se hace la aclaración que en el código se variaba el url_noticias por cada diario y por cada año (todos los días del año), obteniendo así el total de 32411 registros. En algunos años o diarios, debido a restricciones de seguridad no se pudo extraer noticias en dichos diarios o años. El consolidado de todas las noticias considerado en el presente trabajo, tanto para el periodo 2020 a 2021 como para el 2020 a 2024, también fueron almacenados en la plataforma GitHub, ver Referencia I.

2. Preprocesamiento y transformación:

El repositorio se decidió como un almacenamiento local. Este preprocesamiento constituye una fase crítica en el proceso KDD, ya que es el puente entre la recolección bruta de noticias y la posterior minería de datos / minería de texto. Dada la naturaleza heterogénea de las fuentes periodísticas, los datos obtenidos presentaban una serie de retos y la variabilidad en la forma de mencionar ubicaciones o categorías de robos.

Con el preprocesamiento se garantizó que la base de noticias cumpliera con los requisitos mínimos de fiabilidad, completitud y consistencia, asegurando que el análisis posterior estuviera fundamentado en una base de noticias robusta y metodológicamente sólida. Se aplicaron técnicas de limpieza semántica básica, que incluyeron la eliminación de stopwords (artículos, preposiciones, conjunciones). Así también, se apoyó en el diccionario de variantes para mapear expresiones como “asalto”, “asaltos” o “asaltante”, todas asociadas a “robo”. Este procedimiento preparó para el conteo sistemático de ocurrencias de palabras y la clasificación posterior por distrito.

En cada registro se distinguió por título, resumen y fecha de la noticia, lo que permitió conservar la estructura para el análisis comparativo entre los distintos niveles. Es por tanto, esta fase el objetivo de una estandarización, normalización y aplicación de georreferenciación; que servirán de base para la minería de datos / minería de texto. El código en lenguaje de programación R, tanto para el periodo 2020 a 2021 como para el 2020 a 2024, asociado al preprocesamiento y transformación fueron almacenados en la plataforma GitHub, ver Referencia I.



3. Minería de Datos / Minería de texto:

Esta fase es la de identificación de patrones, para el contexto del presente trabajo, contribuye para la identificación de zonas de mayor incidencia en robos. Para ello, una de las principales herramientas que se usaron fueron el análisis geoespacial, mediante la información de códigos ubigeo de los distritos de Lima; y otras herramientas de apoyo para la evaluación posterior fueron: tabla de frecuencias, matriz de contingencias, dendogramas, y nubes de palabras. El código en lenguaje de programación R, tanto para el periodo 2020 a 2021 como para el 2020 a 2024, para esta fase también fueron almacenados en la plataforma GitHub, ver Referencia I.

4. Evaluación:

La fase de Evaluación actúa como un puente entre el análisis técnico y la toma de decisiones, garantizando que los patrones identificados sobre robos en Lima no solo sean estadísticamente correctos, sino también socialmente relevantes y útiles. Es así que, el objetivo de esta evaluación es evaluar la validez, coherencia, interoperabilidad y utilidad práctica de los patrones identificados mediante técnicas de minería de datos y análisis geoespacial, con el fin de asegurar que el conocimiento obtenido represente adecuadamente la dinámica de los robos en Lima y sea relevante para la toma de decisiones en seguridad ciudadana. Para ello, será parte de la evaluación responder las siguientes preguntas:

- ¿Los patrones de robos identificados son consistentes a lo largo del período analizado y no responden a eventos aislados o aleatorios?
- ¿Existe coherencia entre los resultados obtenidos mediante tablas de frecuencia, matrices de contingencia y análisis geoespacial?
- ¿Los distritos identificados como críticos presentan una lógica espacial coherente con la estructura urbana y socioeconómica de Lima?
- ¿Los agrupamientos de distritos obtenidos mediante dendrogramas son interpretables y socialmente coherentes?
- ¿Los resultados pueden ser comprendidos y utilizados por autoridades locales y actores no especializados en análisis de datos?

5. Presentación del conocimiento:

El objetivo de esta fase es presentar de forma clara, comprensible y orientada a la toma de decisiones el conocimiento validado sobre los patrones de robos en Lima, mediante visualizaciones y resúmenes interpretativos que permitan su aplicación práctica en la gestión de la seguridad ciudadana. Los resultados se presentan utilizando mapas temáticos, gráficos estadísticos y resúmenes interpretativos, permitiendo identificar distritos críticos, modalidades de robo predominantes y patrones territoriales de manera intuitiva.

Esta etapa se subdivide a su vez en las siguientes etapas listadas a continuación:

1. Interpretación de patrones: Consiste en analizar el significado de los resultados obtenidos mediante minería de datos y determinar su relevancia en el contexto del problema. Aquí se traduce el hallazgo técnico en una explicación comprensible.
2. Evaluación del conocimiento descubierto: Implica validar que los patrones encontrados sean: correctos (válidos), novedosos, útiles, no producto del azar. Esta evaluación puede incluir métricas estadísticas y juicio experto.
3. Visualización del conocimiento: Es una etapa esencial en la cual los patrones se presentan mediante: gráficos, mapas, dendrogramas, representaciones geoespaciales. Han et al. (2012) destacan que la visualización facilita la comprensión humana del conocimiento extraído.
4. Presentación y comunicación al usuario final: El conocimiento debe organizarse en reportes o productos interpretables para apoyar la toma de decisiones. Esto puede incluir: informes analíticos, mapas de riesgo, modelos predictivos explicados.
5. Integración del conocimiento en procesos de decisión: Finalmente, el conocimiento descubierto puede ser incorporado en sistemas reales, políticas públicas o modelos de monitoreo. En criminología urbana, esto puede traducirse en alertas, focalización territorial o estrategias de prevención.



Figura 3.1: Etapas de la presentación del conocimiento

3.2 Descripción de los datos

La data por analizar viene de cuatro diarios de relevancia nacional, y de cinco años, del 2020 al 2024, con el objetivo de obtener una mayor trazabilidad. Toda esta data fue extraída de la página web de cada diario usando códigos de programación en R, y delimitando la extracción principalmente en los atributos de la noticia: Título, Resumen, Fecha y Hora.

Cuadro 3.1: Resumen final: Cantidad de noticias por año y por diario (2020–2024)

Año	Diario	Cantidad de noticias analizadas
2020	Correo	1040
2020	El Comercio	5999
2020	Gestión	0
2020	Ojo	19
2021	Correo	2320
2021	El Comercio	583
2021	Gestión	50
2021	Ojo	2
2022	Correo	2762
2022	El Comercio	73
2022	Gestión	677
2022	Ojo	11
2023	Correo	249
2023	El Comercio	2860
2023	Gestión	880
2023	Ojo	977
2024	Correo	5804
2024	El Comercio	5600
2024	Gestión	1500
2024	Ojo	1005
		32411

Capítulo 4: Hallazgos encontrados

4.1. Extracción de datos

Los datos extraídos corresponden a diarios nacionales de relevancia en el Perú, los cuales son: El Comercio, Gestión, Correo y Ojo. Estos registros fueron entre los años 2020 al 2024, a partir de una extracción de data de la página web de cada diario usando códigos de programación en R (proceso web scraping); se anexa al presente trabajo la data extraída. Como se indicó anteriormente, el consolidado de noticias extraídas y el código web scraping para obtener dicho consolidado fueron almacenados en la plataforma GitHub, ver Referencia I.

4.2. Análisis de los datos

4.2.1 Descripción

El global de la data analizada fue de 4 años, entre los años 2020 a 2024. La distribución de la data se muestra a continuación (obtenido del código de programación establecido), en relación a las palabras relacionadas a robo, así como la distribución en relación a los distritos de Lima metropolitana (primeros 43 distritos) y Callao (los últimos 7 distritos).

Cuadro 4.1: Frecuencia de palabras asociadas a Robo - Período 2020 a 2024

ID	Delito	Frecuencia
1	Asalto	69
2	Asaltos	15
3	Asaltante	3
4	Asaltantes	11
5	Asaltar	22
6	Asaltaron	12
7	Asalta	0
8	Delincuencia	51
9	Delincuente	59
10	Delincuentes	201
11	Delinquir	3
12	Robos	45
13	Robo	129
14	Robaba	6
15	Robaban	7

16	Robaron	23
17	Rateros	1
18	Ratero	0
19	Roba	24
20	Robar	71
21	Ladrones	44
22	Ladrón	23
23	Raquetero	0
24	Raqueteros	0
25	Latrocínio	0
26	Despojo	0
27	Rapiña	0
28	Saqueo	3
29	Atracos	4
30	Atraco	11
31	Expropiación	3
32	Extorsión	44
33	Hurto	17
34	Sustracción	3
35	Estafas	12
36	Estafadores	5
37	Estafa	27
38	Estafan	3
39	Hampones	42
40	Hampón	8
41	Sustrae	0
42	Sustraer	9
43	Maleante	0
44	Maleantes	15

Cuadro 4.2: Frecuencia de robos por Distritos - Período 2020 a 2024

ID	Distrito	Frecuencia
1	Ancón	17
2	Ate	62
3	Barranco	45
4	Breña	23
5	Carabayllo	25
6	Chaclacayo	6
7	Chorrillos	69
8	Cieneguilla	5
9	Comas	49
10	El Agustino	23
11	Independencia	82
12	Jesús María	33
13	La Molina	36
14	La Victoria	195
15	Cercado de Lima	60
16	Lince	21
17	Los Olivos	46
18	Chosica	10
19	Lurín	19
20	Magdalena del Mar	6
21	Miraflores	112
22	Pachacámac	3
23	Pucusana	2
24	Pueblo Libre	21
25	Puente Piedra	51
26	Punta Hermosa	9
27	Punta Negra	1
28	Rímac	35
29	San Bartolo	1
30	San Borja	37
31	San Isidro	35

32	San Juan de Lurigancho	72
33	San Juan de Miraflores	19
34	San Luis	25
35	San Martín de Porres	40
36	San Miguel	73
37	Santa Anita	33
38	Santa María del Mar	0
39	Santa Rosa	66
40	Santiago de Surco	2
41	Surquillo	11
42	Villa el Salvador	28
43	Villa María del Triunfo	1
44	Callao	201
45	Ventanilla	32
46	Carmen de la Legua Reynoso	0
47	Bellavista	11
48	La Perla	6
49	Mi Perú	1
50	La Punta	11

Por otro lado, para el caso del periodo 2020 a 2021, años en que ocurrió la pandemia del Coronavirus en el Perú, la distribución de la data en específico es como es muestra a continuación:

Cuadro 4.3: Frecuencia de palabras asociadas a Robo - Período 2020 a 2021

ID	Delito	Frecuencia
1	Asalto	69
2	Asaltos	15
3	Asaltante	3
4	Asaltantes	11
5	Asaltar	22
6	Asaltaron	12
7	Asalta	0

8	Delincuencia	51
9	Delincuente	59
10	Delincuentes	201
11	Delinquir	3
12	Robos	45
13	Robo	129
14	Robaba	6
15	Robaban	7
16	Robaron	23
17	Rateros	1
18	Ratero	0
19	Roba	24
20	Robar	71
21	Ladrones	44
22	Ladrón	23
23	Raquetero	0
24	Raqueteros	0
25	Latrocínio	0
26	Despojo	0
27	Rapiña	0
28	Saqueo	3
29	Atracos	4
30	Atraco	11
31	Expropiación	3
32	Extorsión	44
33	Hurto	17
34	Sustracción	3
35	Estafas	12
36	Estafadores	5
37	Estafa	27
38	Estafan	3
39	Hampones	42
40	Hampón	8

41	Sustrae	0
42	Sustraer	9
43	Maleante	0
44	Maleantes	15

Cuadro 4.4: Frecuencia de robos por Distritos - Período 2020 a 2021

ID	Distrito	Frecuencia
1	Ancón	17
2	Ate	62
3	Barranco	45
4	Breña	23
5	Carabayllo	25
6	Chaclacayo	6
7	Chorrillos	69
8	Cieneguilla	5
9	Comas	49
10	El Agustino	23
11	Independencia	82
12	Jesús María	33
13	La Molina	36
14	La Victoria	195
15	Cercado de Lima	60
16	Lince	21
17	Los Olivos	46
18	Chosica	10
19	Lurín	19
20	Magdalena del Mar	6
21	Miraflores	112
22	Pachacámac	3
23	Pucusana	2
24	Pueblo Libre	21
25	Puente Piedra	51

26	Punta Hermosa	9
27	Punta Negra	1
28	Rímac	35
29	San Bartolo	1
30	San Borja	37
31	San Isidro	35
32	San Juan de Lurigancho	72
33	San Juan de Miraflores	19
34	San Luis	25
35	San Martín de Porres	40
36	San Miguel	73
37	Santa Anita	33
38	Santa María del Mar	0
39	Santa Rosa	66
40	Santiago de Surco	2
41	Surquillo	11
42	Villa el Salvador	28
43	Villa María del Triunfo	1
44	Callao	201
45	Ventanilla	32
46	Carmen de la Legua Reynoso	0
47	Bellavista	11
48	La Perla	6
49	Mi Perú	1
50	La Punta	11

Así también, como parte del análisis, para una mejor visualización de la distribución de las palabras asociadas a robo, tanto en el periodo 2020 a 2024 como para el periodo de pandemia (2020 a 2021), se hizo uso de la herramienta Nube de palabras. A continuación, se muestra los resultados respectivos (figura 4.1 y figura 4.2):

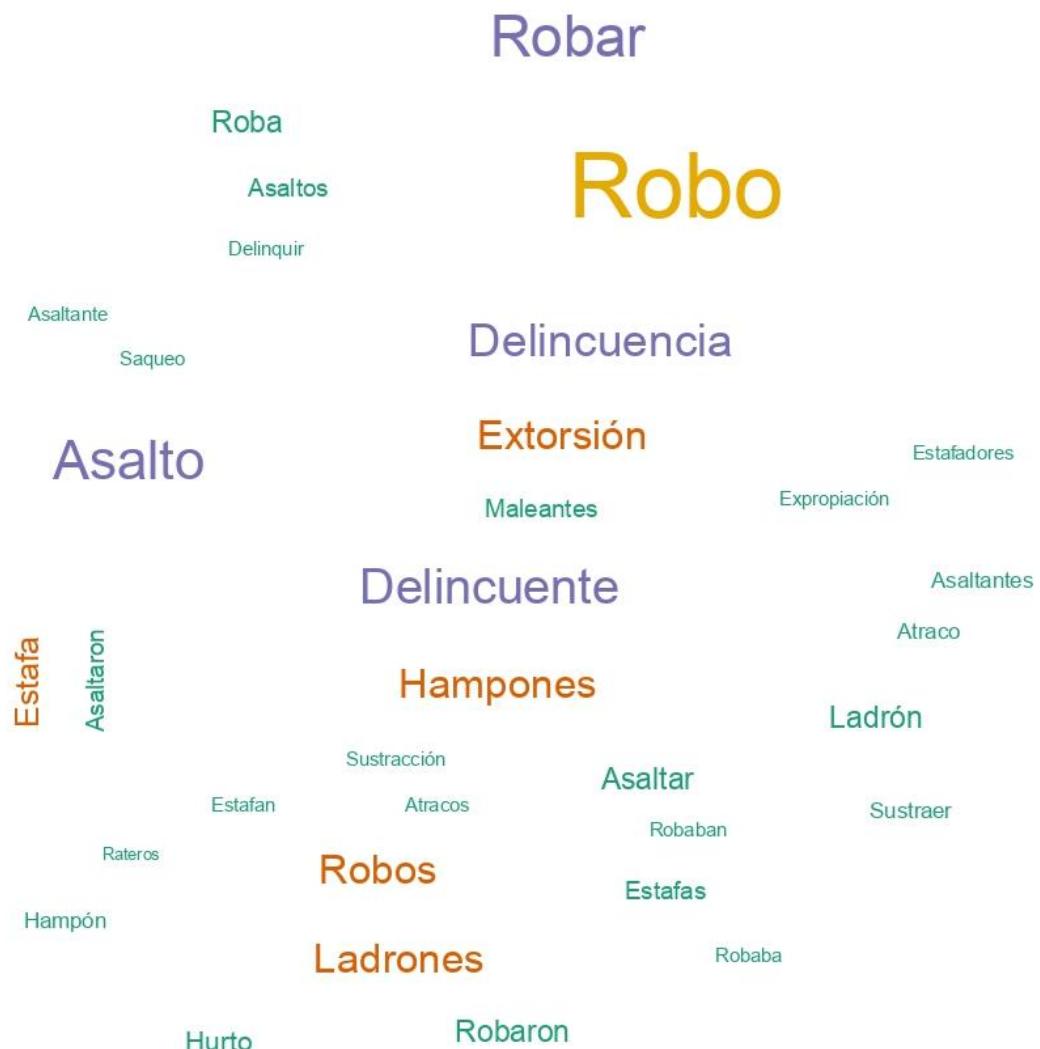


Figura 4.1: Nube de palabras asociadas a Robo – periodo 2020 al 2024



Figura 4.2: Nube de palabras asociadas a Robo – periodo 2020 al 2021

Como se aprecia al comparar ambas figuras, los resultados en ambos períodos son semejantes manteniéndose palabras “Robo”, “Robar”, “Delincuente” en ambos períodos, debido a ello no ameritaría un análisis focalizado para cada período.

Asimismo, también se realizó una Nube de palabras, pero considerando los distritos de Lima metropolitana y del Callao, es decir los 50 distritos. Los resultados se muestran a continuación (figura 4.3 y figura 4.4):

Miraflores

Magdalena del Mar

Santa Rosa

Rímac

Chorrillos

Puente Piedra

Ancón

San Miguel

Lurín

San Bartolo

Ate

La Punta

Chaclacayo

Lince

Cieneguilla

Punta Negra

Independencia

La Perla

Ventanilla

Surquillo

La Molina

Pachacámac

Pucusana

Mi Perú
Villa María del Triunfo

Breña

San Luis

Chosica

El Agustino

Punta Hermosa

Bellavista

San Juan de Miraflores

Comas

Pueblo Libre

San Martín de Porres

Santiago de Surco

Los Olivos

Carabayllo

Figura 4.3: Nube de palabras asociadas a Distritos – periodo 2020 al 2024



Figura 4.4: Nube de palabras asociadas a Distritos – periodo 2020 al 2021

De igual modo, como en la comparación anterior, las nubes de palabras de distritos son semejantes; siendo los distritos de mayor incidencia de robo “Independencia”, “Comas”, “Miraflores”, entre otros; y no correspondería un análisis focalizado para cada periodo.

Asimismo, se hace mención como parte del análisis que se realizó una matriz de contingencia para ambos períodos (2020 a 2024) y (2020 a 2021). Ambas matrices fueron almacenadas en la plataforma GitHub, ver Referencia I. Ello con la finalidad de organizar la distribución conjunta entre modalidades léxicas asociadas al robo y distritos de Lima, permitiendo cuantificar la frecuencia de co-ocurrencia entre ambas variables.

categóricas. Esta estructura constituye el insumo fundamental para el análisis de correspondencias, ya que posibilita explorar asociaciones territoriales y patrones espaciales a partir de datos textuales previamente georreferenciados.

4.2.2 Clasificación

Para la clasificación, me apoyé en dendogramas, los cuales muestran la cantidad de cluster tanto para el periodo de los años 2020 a 2024, y para el periodo de pandemia 2020 a 2021. Los dendogramas respectivos son (ver figura 4.5 y figura 4.6):

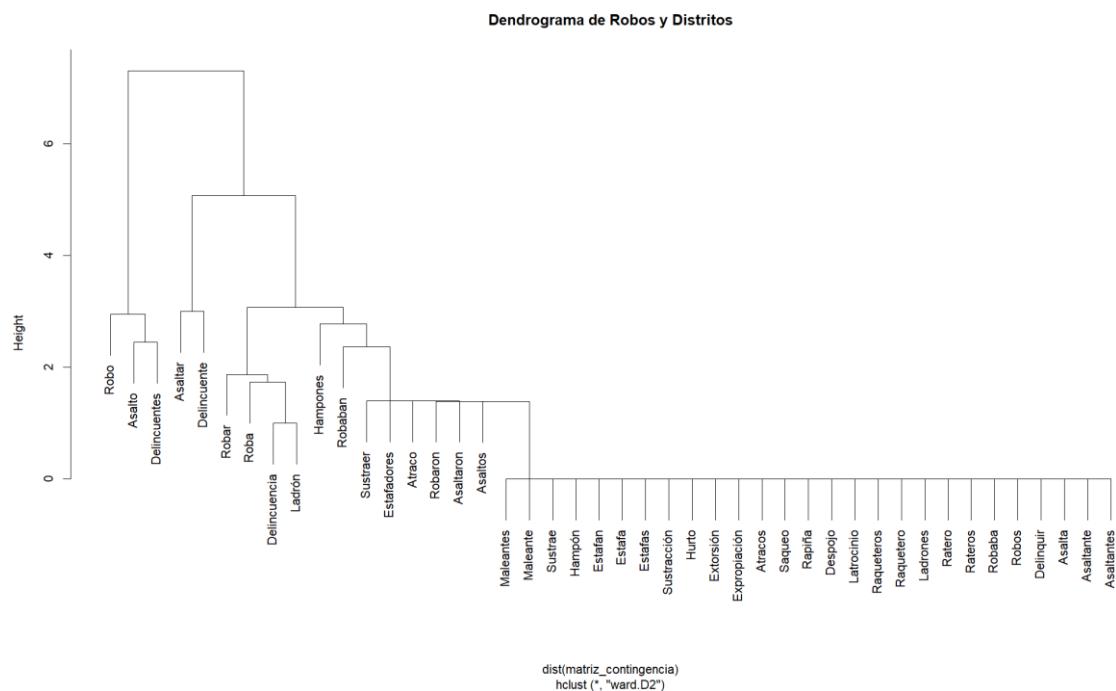


Figura 4.5: Dendrograma de palabras asociadas a Robo – periodo 2020 a 2024

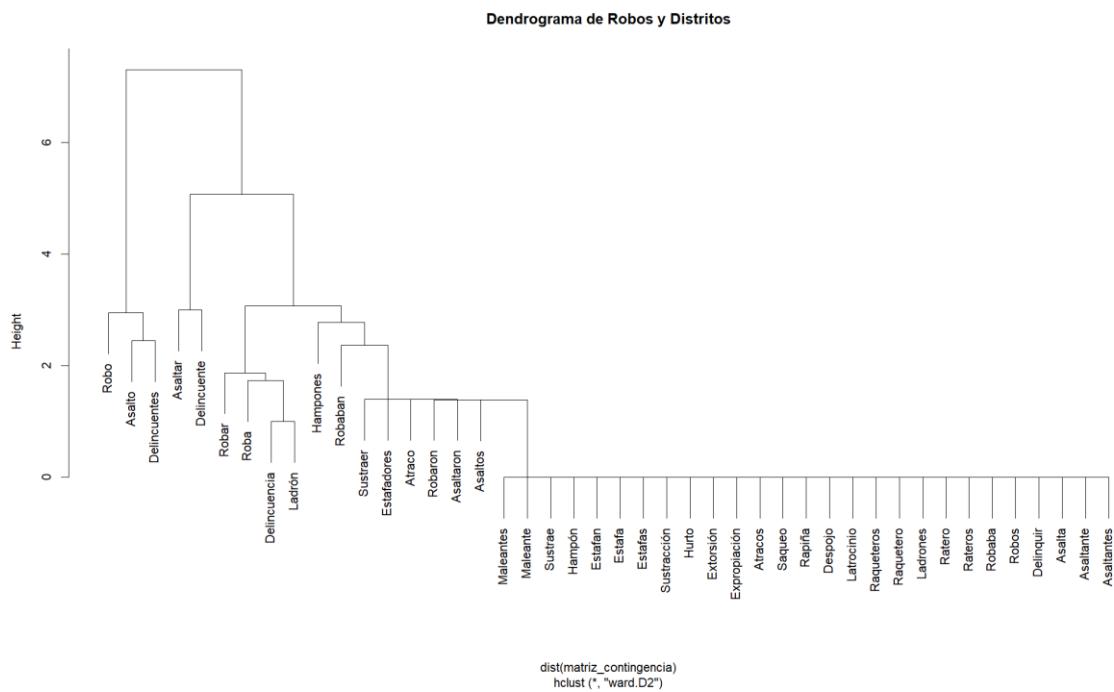


Figura 4.6: Dendrograma de palabras asociadas a Robo – periodo 2020 a 2021

Entre ambos dendogramas, se puede apreciar semejanza. Sin embargo, el valor agregado obtenido es poder reducir la cantidad de palabras planteadas asociadas a robo, de esta manera la búsqueda de noticias asociadas a ello sería más eficiente.

Debido a que estos dendogramas no mostraron gran valor agregado fue necesario la realización de dendogramas asociados a los distritos. Los resultados de ello se muestran a continuación (ver figura 4.7 y figura 4.8):

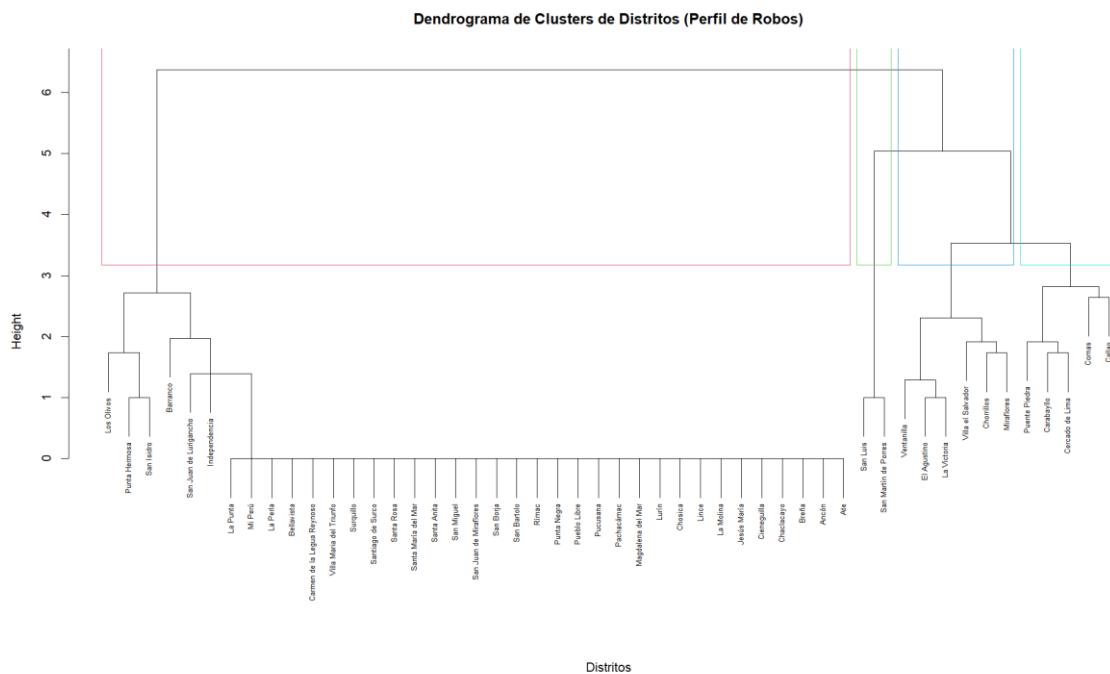


Figura 4.7: Dendrograma de palabras asociadas a los Distritos – periodo 2020 a 2024

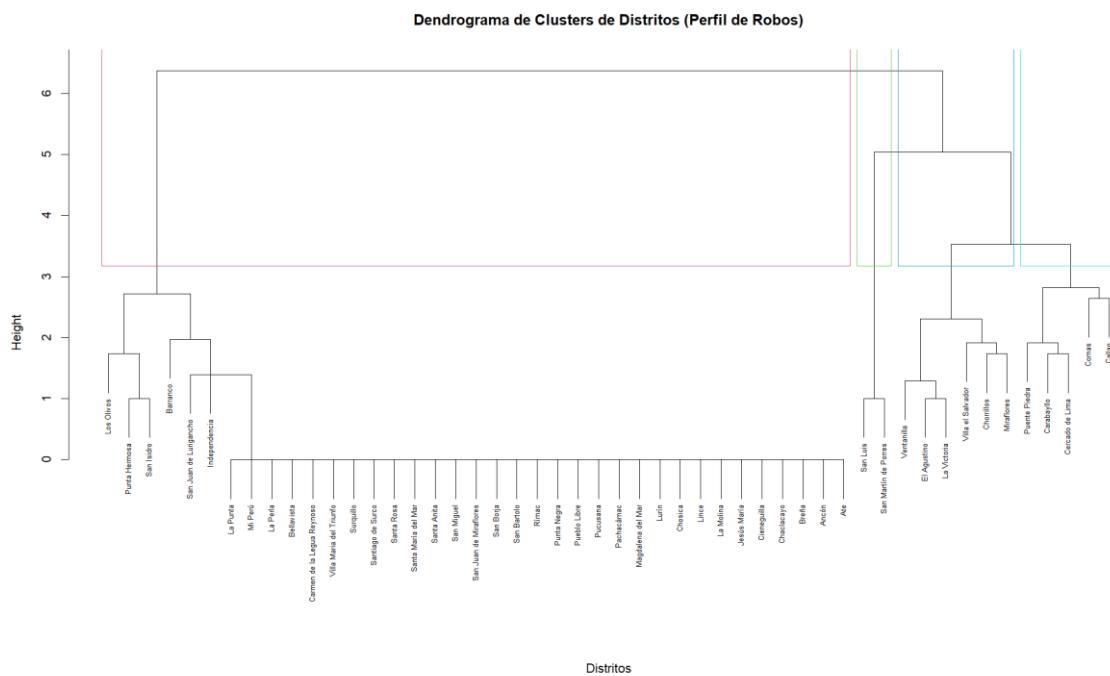


Figura 4.8: Dendrograma de palabras asociadas a los Distritos – periodo 2020 a 2021



Los resultados de los dendogramas de ambos períodos no muestran diferencia; sin embargo; es de valor su uso para aplicación de políticas públicas focalizadas por zonas distritales y así realizar un mejor uso de los recursos públicos.

Corresponde ahora responder las preguntas de la fase de Evaluación, planteadas en la sección anterior, las cuales son:

¿Los patrones de robos identificados son consistentes a lo largo del período analizado y no responden a eventos aislados o aleatorios?

Como se demostró en las nubes de palabras y dendogramas, los resultados fueron similares, independientemente del período que vengan, es decir sea del 2020 a 2024 o del período 2020 a 2021 (período pandemia). Por tanto, no existen eventos aislados durante el análisis.

¿Existe coherencia entre los resultados obtenidos mediante tablas de frecuencia, matrices de contingencia y análisis geoespacial?

Sí, por ejemplo, los resultados en frecuencia, nube de palabras y mapa muestran que los distritos con mayor incidencia en robos son: “Independencia”, “La Victoria”, “Miraflores”.

¿Los distritos identificados como críticos presentan una lógica espacial coherente con la estructura urbana y socioeconómica de Lima?

No necesariamente, por ejemplo en el caso del distrito de Miraflores, es un distrito con alta percepción de seguridad; sin embargo como muestras los resultados la frecuencia de robos es mayor que incluso distritos de los que en Lima se les dice “los conos”.

¿Los agrupamientos de distritos obtenidos mediante dendrogramas son interpretables y socialmente coherentes?

Sí, pues se aprecia como los dendrogramas por distritos se han agrupado sobre la base de la frecuencia de robos, lo muestra claramente una clasificación.

¿Los resultados pueden ser comprendidos y utilizados por autoridades locales y actores no especializados en análisis de datos?

Sí, los resultados podrían ser usados para asignación de recursos públicos y diseño de políticas públicas basadas en evidencia.

4.2.3 Análisis Espacial

El análisis por georreferenciación se realizó mediante la obtención de un mapa nacional de Lima, y ver cómo se distribuyen las noticias en los distritos. Para ello también se usó códigos de programación en R, solo considerando los 43 distritos de Lima metropolitana.

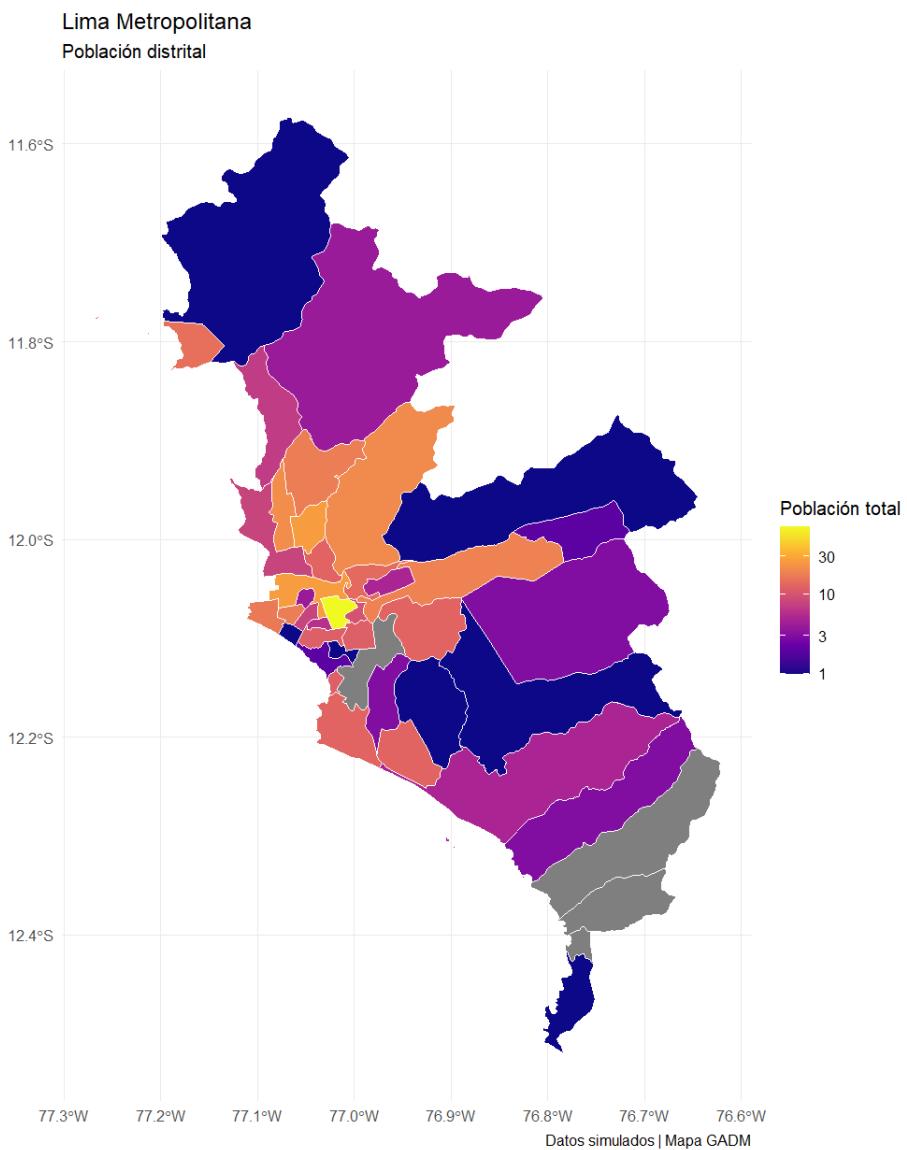


Figura 4.9: Mapa de robos – periodo 2020 a 2024

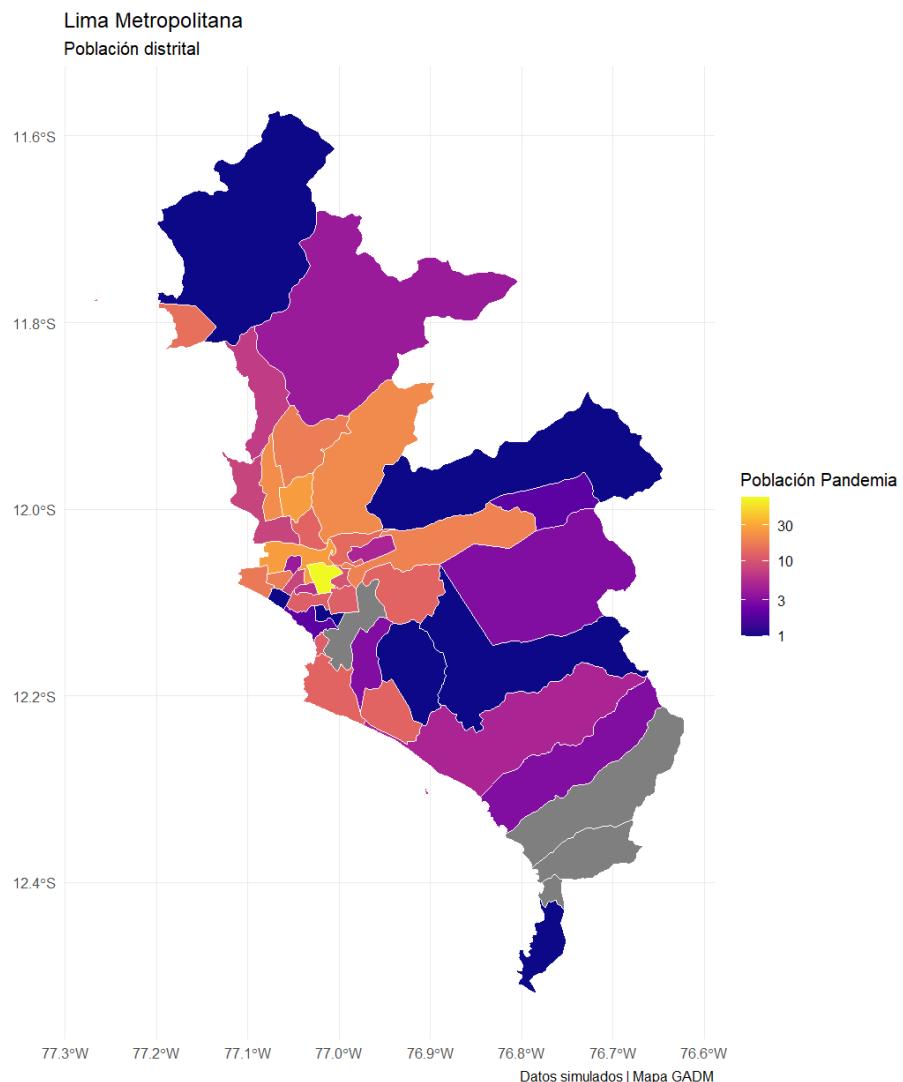


Figura 4.10: Mapa de robos – periodo 2020 a 2021

Estos mapas presentados muestran similitud, y principalmente su uso es para una mejor visualización o presentación de los resultados, y comparativo con otras ciudades o regiones.



Capítulo 5: Conclusiones

5.1. Conclusiones del trabajo

5.2. Limitaciones técnicas y oportunidades de mejora

5.3. Trabajos a futuro



Capítulo 6: Referencias

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77–84. <https://doi.org/10.1145/2133806.2133826>

Chainey, S., Tompson, L., & Uhlig, S. (2008). The utility of hotspot mapping for predicting spatial patterns of crime. *Security Journal*, 21(1–2), 4–28.
<https://doi.org/10.1057/palgrave.sj.8350066>

Chen, H., Chung, W., & Xu, J. (2004). Crime data mining: A general framework and some examples. *Computer*, 37(4), 50–56. <https://doi.org/10.1109/MC.2004.1297301>

Espinoza-Ramírez, A., Nakano, M., Sánchez-Pérez, G., & Arista-Jalife, A. (2018). Sistemas de información geográfica y su análisis aplicado en zonas de delincuencia en la Ciudad de México. *Información Tecnológica*, 29(5), 235–246.
<https://doi.org/10.4067/S0718-07642018000500235>

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis* (5th ed.). Wiley.

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3), 37–54.
<https://doi.org/10.1609/aimag.v17i3.1230>

Freedman, D., Pisani, R., & Purves, R. (2007). *Statistics* (4th ed.). W. W. Norton & Company.

Greenacre, M. (2017). *Correspondence analysis in practice* (3rd ed.). CRC Press.

Hackeloeer, A., Klasing, K., Krisp, J. M., & Meng, L. (2014). Georeferencing: A review of methods and applications. *International Journal of Geographical Information Science*, 28(8), 1501–1528. <https://doi.org/10.1080/13658816.2014.883098>

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Harries, K. (2001). *Crime mapping and spatial analysis*. National Institute of Justice.

Hearst, M. A. (1999). Untangling text data mining. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics* (pp. 3–10). Association for Computational Linguistics. <https://doi.org/10.3115/1034678.1034679>



Hernández Salazar, L. (2020). *Métricas multivariantes textuales georreferenciadas: Caso aplicado al Estado de Veracruz 2020* [Tesis de maestría, Universidad Autónoma de Guerrero, Facultad de Matemáticas, Maestría en Métodos Estadísticos Aplicados].

Hill, M. (2006). *Geographic information systems: Concepts and applications*. CRC Press.

International Organization for Standardization. (2003). *ISO 19112:2003 geographic information — Spatial referencing by geographic identifiers*. ISO.

International Organization for Standardization. (2007). *ISO 19111:2007 geographic information — Spatial referencing by coordinates*. ISO.

Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing* (3rd ed., draft). <https://web.stanford.edu/~jurafsky/slp3/>

Kantardzic, M. (2011). *Data mining: Concepts, models, methods, and algorithms* (2nd ed.). Springer.

Kinne, J. (2018). Exploring the use of text mining for innovation indicator development. *Technology Innovation Management Review*, 8(6), 43–49.
<https://doi.org/10.22215/timreview/1160>

Leidner, J. L. (2008). *Toponym resolution in text: Annotation, evaluation and applications of spatial grounding of place names*. Springer.

Liu, B. (2012). *Sentiment analysis and opinion mining*. Morgan & Claypool Publishers.
<https://doi.org/10.2200/S00416ED1V01Y201204HLT016>

Liu, X., Andris, C., & Ratti, C. (2010). Uncovering cabdrivers' behavior patterns from their digital traces. *Computers, Environment and Urban Systems*, 34(6), 541–548.
<https://doi.org/10.1016/j.compenvurbsys.2010.07.003>

Mitchell, R. (2018). *Web scraping with Python: Collecting more data from the modern web* (2nd ed.). O'Reilly Media.

Moreno Jiménez, A., & Grijalva Eternod, A. E. (2022). Alternativas para la medición de la delincuencia urbana y la identificación de zonas criminógenas: Nuevos indicadores basados en la presencia de población. *Estudios Geográficos*, 83(293), e121.
<https://doi.org/10.3989/estgeogr.2022127.127>

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1), 3–26.
<https://doi.org/10.1075/li.30.1.03nad>



Perú. Decreto Legislativo N.^º 635, Código Penal. (1991). *Diario Oficial El Peruano*.
<https://www.gob.pe/institucion/indecopi/normas-legales/5582266-635>

Project Guru. (s. f.). *Text mining: Analyzing unstructured data*.
<https://www.projectguru.in/text-mining-analyzing-unstructured-data/>

Real Academia Española. (s. f.). *Robo*. En *Diccionario de la lengua española* (23.^a ed.). <https://dle.rae.es/robo>

Ricci, F., Rokach, L., & Shapira, B. (Eds.). (2011). *Recommender systems handbook*. Springer. <https://doi.org/10.1007/978-0-387-85820-3>

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), 513–523.
[https://doi.org/10.1016/0306-4573\(88\)90021-0](https://doi.org/10.1016/0306-4573(88)90021-0)

Sarawagi, S. (2008). Information extraction. *Foundations and Trends® in Databases*, 1(3), 261–377. <https://doi.org/10.1561/1900000003>

Tao, L., Xie, Z., Xu, D., Ma, K., Qiu, Q., Pan, S., & Huang, B. (2022). Geographic named entity recognition by employing natural language processing and an improved BERT model. *ISPRS International Journal of Geo-Information*, 11(12), 598.
<https://doi.org/10.3390/ijgi11120598>

Villa Monte, A. (2023). *Minería de datos: Tema 2* [Diapositivas en PDF]. Universidad Internacional de Valencia.

Wang, X., Gerber, M. S., & Brown, D. E. (2012). Automatic crime prediction using events extracted from Twitter posts. In *Social Computing, Behavioral-Cultural Modeling and Prediction* (pp. 231–238). Springer. https://doi.org/10.1007/978-3-642-29047-3_28



Capítulo 7: Anexos

Anexo I: Repositorio de código y archivos

Link Github:

<https://github.com/AlexanderFernandezE/TFM-VIU-Alexander-Fernandez>



Anexo II: Trazabilidad de uso de IA

Herramienta IA	Mes de consulta	Sección	Prompts Utilizados
ChatGPT	Febrero 2026	Estado del arte	En una referencia por ejemplo al colocar: Fayyad et al. (1996) ¿Qué significa et al.?
ChatGPT	Febrero 2026	Estado del arte	En relación a la elaboración de una TFM llamada "Minería de textos georreferenciados aplicada al contexto de robos en Lima, Perú" Citar referencias papers asociadas a la TFM
ChatGPT	Febrero 2026	Estado del arte	En el texto de la TFM, al colocar: Adaptado de ... ¿cómo debo colocarlo?
ChatGPT	Febrero 2026	Hallazgos encontrados	Si lo adjunto es una base de datos de noticias, dame una tabla resumen de cantidad de noticias por año y por diario.
ChatGPT	Febrero 2026	Referencias	En relación a bibliografía en una TFM, ordena este listado como debería estar correctamente, y también coloca en cursiva si corresponde.