Mitigación de problemas de seguridad ciudadana con la aplicación de técnicas de Machine Learning en ciudades inteligentes

Jhonatan Felipe Sossa Rojo

Informe final de trabajo de grado para optar al título de Ingeniero de telecomunicaciones

Tutor
Luis Alejandro Fletscher Bocanegra, doctor en ingeniería.

Universidad de Antioquia
Facultad de Ingeniería
Ingeniería de Telecomunicaciones
Medellín
2022

1. Resumen

La seguridad ciudadana es un problema que preocupa a los gobiernos de las grandes y pequeñas ciudades y al que se destina una cantidad significativa del presupuesto de las naciones. Por este motivo, a través de modelos y herramientas tecnológicas, se busca actualmente enfrentar el problema y reforzar la seguridad ciudadana para mejorar la calidad de vida de las personas que habitan los territorios.

Este proyecto consiste en la aplicación de técnicas de aprendizaje de máquina a la predicción de incidentes delictivos que tienen alta probabilidad de suceder en un espacio y tiempos debidamente delimitados, utilizando para ello características extraídas de tweets. Con esto, se busca darle a las autoridades una herramienta que permita planear adecuadamente las labores de vigilancia en las ciudades. Además, la predicción y análisis de incidentes delictivos en zonas específicas permite a los gobiernos tomar acciones preventivas y correctivas y también entender los territorios de una manera más acertada.

2. Introducción

En el marco del proyecto "ADMINISTRACIÓN INTELIGENTE DE PROBLEMAS DE SEGURIDAD CIUDADANA A TRAVÉS DE MODELOS Y HERRAMIENTAS GENERADOS A PARTIR DE PLATAFORMAS PARA TERRITORIOS INTELIGENTES APOYADAS POR ESTRATEGIAS DE PARTICIPACIÓN CIUDADANA EN LA CIUDAD DE MEDELLÍN. Código SIGP 67040", el propósito del componente dos es aplicar técnicas de inteligencia computacional sobre los datos para solucionar problemáticas de seguridad ciudadana. Con este trabajo de grado, se busca sentar las primeras bases en el proyecto sobre la aplicación de técnicas de machine learning sobre datos capturados en ciudades inteligentes. Para ello, se planteó realizar una búsqueda de bases de datos que se han recolectado en ciudades inteligentes y han sido usadas para la aplicación de técnicas de machine learning en soluciones de seguridad ciudadana, para posteriormente, escoger una de ellas que contenga información valiosa y con la cual se pueda resolver un problema de seguridad que sea aplicable al Área Metropolitana. Una vez definida la base de datos a utilizar, se

justificará y aplicará alguna técnica de machine learning para resolver el problema y evaluar su desempeño con las métricas escogidas. Estos resultados le brindarán al componente de analítica información valiosa sobre datos relevantes para capturar en el municipio donde se decida implementar el proyecto.

3. Objetivos

3.1. Objetivo general:

Evaluar técnicas de machine learning aplicadas a la solución de problemas de seguridad ciudadana en el contexto de ciudades inteligentes.

3.2. Objetivos específicos:

1. Recolectar bases de datos obtenidas en ciudades inteligentes y seleccionar aquella que contenga información útil para resolver problemas de seguridad aplicables a los municipios del área metropolitana.
2. Definir la técnica de machine learning a utilizar a partir de la base de datos seleccionada y del problema de seguridad que se desea trabajar
3. Aplicar la técnica de machine learning escogida sobre la base de datos escogida
4. Definir y aplicar las métricas de evaluación del desempeño para el algoritmo de machine learning aplicado sobre la base de datos.

4. Estado del arte

Para comenzar a trabajar sobre seguridad ciudadana en ciudades inteligentes, primero se debe tener un acercamiento a la definición de lo que es una ciudad inteligente. La siguiente, aunque no es una definición única, se acerca bastante al tipo de ciudad inteligente que se quiere lograr a través de este trabajo de grado y, en general, del proyecto en el cual está enmarcado: "Una ciudad es inteligente cuando las inversiones en capital humano y social y la infraestructura en

comunicación tradicional (transporte) y moderna (TIC) impulsan el crecimiento económico sostenible y una alta calidad de vida, con una gestión inteligente de los recursos naturales, a través de una gobernanza participativa" (Caragliu, A. & Del Bo, C. & Nijkamp, P. 2009).

Como complemento a la definición anterior, dentro de una ciudad inteligente se deben aplicar soluciones inteligentes. Estas soluciones inteligentes deben estar lideradas por las nuevas tecnologías que día a día se van desarrollando. Así, una ciudad inteligente debe tener la capacidad de aprovechar la tecnología que conduce a soluciones 'inteligentes', sus aplicaciones permitirán que las ciudades utilicen las TIC y los datos para mejorar la infraestructura y los servicios. (Srivastava, A. Bisht and N. Narayan. 2017).

Según Latin American and Caribbean Institute for Economic and Social Planning (1997), la seguridad ciudadana se puede entender como la preocupación por la calidad de vida y la dignidad humana en términos de:

- Libertad
- Acceso, y
- Oportunidades sociales

para todos los individuos que comparten un entorno social delimitado por el territorio de un país.

Una tecnología que pretende brindar estas soluciones 'inteligentes' a las ciudades es la inteligencia artificial (IA)". La IA es el intento de una máquina de explicar el comportamiento del sistema (humano) que está tratando de modelar…" Schank, R. C. (1991). Una de las soluciones 'inteligentes' a las que hace referencia uno de los párrafos anteriores se da por el lado de la seguridad ciudadana. En este campo ha habido avances significativos tales como: (S. Srivastava, A. Bisht and N. Narayan. 2017)

- Sistemas de detección de fraude
- Sistemas de detección avanzados
- Detección de drogas

- Computación avanzada e interfaz humano-computadora
- Software de reconocimiento de voz avanzado

A su vez, en este contexto una de las aplicaciones más estudiadas es la detección de disparos, debido a que "La detección de armas juega un rol vital en la gestión de la seguridad, la protección y la vigilancia" (A. Jain, Aishwarya and G. Garg. 2020). El enfoque utilizado en (A. Jain, Aishwarya and G. Garg. 2020) aprovecha la gran cantidad de cámaras de videovigilancia que hay en la actualidad para hacer reconocimiento de disparos utilizando un método llamado Haar Cascade Classifier.

Otra aplicación interesante en el entorno de la seguridad ciudadana usando técnicas de machine learning aborda el tema de utilizar la información abierta que generan entidades gubernamentales a diario. Con esta información se pueden plantear enfoques como el de G. B. Rocca (2016) donde utilizan una de estas bases de datos para encontrar el camino más corto y seguro entre dos puntos determinando qué tipos de 'eventos' hay en la zona. Para resolverlo, utilizan dos algoritmos diferentes:

- El primero de ellos es Multiple Logistic Regression (MLR) y está basado en la regresión logística original, pero extendido para predecir más de dos clases. Una de las ventajas de la regresión logística es que funciona muy bien con grandes cantidades de datos y tiene un gran comportamiento si el problema es linealmente separable. (G. B. Rocca, M. Castillo-Cara, R. A. Levano, J. V. Herrera and L. OrozcoBarbosa. 2016).
- El segundo es Random Forest Algorithm el cual hace uso de árboles de decisión simples los cuales se pueden definir como un proceso para llegar a una respuesta haciendo una serie de preguntas. Random Forest utiliza cada uno de estos árboles para elegir la predicción a partir de la decisión más votada. (G. B. Rocca, M. Castillo-Cara, R. A. Levano, J. V. Herrera and L. OrozcoBarbosa. 2016).

Dentro del amplio espectro de la seguridad ciudadana propiciando la participación ciudadana, se encuentra lo propuesto por Matthew S (2014). Un

sistema de predicción de crímenes utilizando Twitter y KDE (Kernel Density Estimation). En dicho artículo, recolectan una base de datos que recopila los tweets realizados en la ciudad de Chicago entre los meses de enero y marzo del año 2013. Adicional a la base de datos de tweets, también cuentan con una base de datos (extraída del portal oficial de datos de la ciudad de Chicago) donde se consolidan los crímenes cometidos en la ciudad durante el mismo periodo de tiempo. Los crímenes que allí se documentan se clasifican entre 27 tipos diferentes. Con la base de datos de crímenes se extrae una estimación de la densidad del kernel para cada uno de los tipos de crimen y se utiliza como primera característica para el entrenamiento del modelo. Luego de la base de datos de tweets y con apoyo del algoritmo Latent Dirichlet Allocation (LDA), encuentran n características adicionales las cuales, en conjunto con la característica de estimación de la densidad, conforman las características que describen las muestras con las que se entrena el modelo. Luego, a través de una ventana móvil de 30 días para entrenamiento y 1 día para test, predicen la probabilidad de que suceda un tipo específico de crímen en una zona delimitada de la ciudad de Chicago.

Siguiendo con la línea de predicción de crímenes con características extraídas de redes sociales, en Rumi, S. K. (2018) realiza la predicción a partir de características históricas, demográficas, geográficas y dinámicas, donde las características dinámicas son extraídas con ayuda de Foursquare que es un servicio basado en localización web aplicada a las redes sociales.

Finalmente, Helmstetter, S. (2021) utiliza características extraídas de Twitter para clasificar Fake News usando supervisión débil y comparando los resultados obtenidos para diferentes clasificadores binarios.

5. Marco Teórico

5.1. Topic Modeling:

El modelado de temas es un enfoque de aprendizaje no supervisado en el que el modelo identifica los temas mediante la detección de patrones como grupos de palabras y frecuencias. Los resultados de un modelo de temas son:

1. Clusters de documentos que el modelo ha agrupado en base a temas
2. Grupos de palabras (tópicos) que el modelo ha utilizado para inferir las relaciones

Se deben tener en cuenta dos asunciones importantes en el proceso de modelado de temas:

1. El supuesto de distribución indica que temas similares hacen uso de palabras similares
2. El supuesto de mezcla estadística indica que cada documento trata varios temas.

En pocas palabras, para un corpus dado de documentos, cada documento se puede representar como una distribución estadística de un conjunto fijo de temas. El papel del modelado de temas es identificar los temas y representar cada documento como una distribución de estos temas.

5.1.1. Latent Dirichlet Allocation (LDA):

Latent Dirichlet Allocation es un modelo probabilístico generativo de contenido textual que identifica temas coherentes de discusión dentro de colecciones de documentos.

LDA asume que los documentos están compuestos de palabras que ayudan a determinar los temas y asigna los documentos a una lista de temas asignando cada palabra en el documento a diferentes temas. Dicha asignación se hace en términos de estimaciones de probabilidad condicional. La probabilidad de que la palabra w_j pertenezca al tema t_k. LDA ignora el orden de aparición de las palabras y la información sintáctica. Los documentos son tratados como una colección de palabras o una bolsa de palabras.

Aunque LDA es un algoritmo no supervisado, al igual que K-means para ejercicios de clasificación no supervisada, se debe indicar cuál es el número de tópicos deseados. Si, por ejemplo, se desean tres tópicos, cada documento puede ser representado como se muestra a continuación:

$$D_i = w_{1i} * Topic_1 + w_{2i} * Topic_2 + w_{3i} * Topic_3 \qquad eq.\ (1)$$

Donde w_{ji} son pesos para cada uno de los tópicos que componen un documento. (By Great Learning Team -. (2020, October 16))

5.2. Selección de características:

Usualmente, algunas de las características de una base de datos no aportan información relevante para la clasificación. Esto porque simplemente son características sin importancia (nombres, identificación, etc), están repetidas o altamente correlacionadas o no aportan información al problema específico.

Debido a esto, es necesario realizar una selección de las características más relevantes para el problema que se está enfrentando.

5.2.1. Análisis de Correlación:

La correlación es un término estadístico que en el uso común se refiere a qué tan cerca están dos variables de tener una relación lineal entre sí. Dos variables que son linealmente dependientes tendrán una correlación más alta que dos variables que no son linealmente dependientes.

Las características con alta correlación son más linealmente dependientes y, por lo tanto, tienen casi el mismo efecto sobre la variable dependiente. Entonces, cuando dos características tienen una alta correlación, se puede eliminar una de ellas. (R, V. (2018, September 11))

5.3. Clasificación:

La clasificación se resume en entrenar un sistema de reconocimiento de patrones que sea capaz de discriminar entre dos o más clases. De esta manera, cuando

nuevos datos sean ingresados al sistema, este debe decidir a qué clase pertenece el nuevo dato.

5.3.1. Máquina de Soporte Vectorial (SVM):

El objetivo de una SVM es encontrar un umbral de decisión que produzca la máxima separación entre las clases. Para esto, usa un concepto de vectores de soporte, los cuales son un subconjunto de observaciones de entrenamiento que están localizados cerca al margen de decisión. Estos vectores de soporte influyen tanto en la localización como en la orientación del margen de decisión. Cuando una SVM está entrenada, una nueva muestra es atribuida a alguna de las clases de acuerdo con cuál lado del hiperplano de separación esta se encuentre (C. M. Bishop. 2006).

5.4. Evaluación de rendimiento:

Los métodos de evaluación de rendimiento utilizado son los siguientes:

5.4.1. Matriz de Confusión:

Es una técnica que resume el rendimiento de un problema de clasificación de aprendizaje automático donde la salida puede ser de dos o más clases. El número de predicciones correctas o incorrectas se resumen con valores de conteo y se desglosa por cada clase. La matriz de confusión 'muestra' las formas en que el modelo de clasificación se confunde cuando hace predicciones.

A partir de la matriz de confusión se pueden extraer varias métricas de performance, entre ellas están: precisión, sensibilidad, especificidad, recall y precisión.

Una matriz de confusión para sistemas de clasificación biclase se muestra en la **Tabla 1**

Tabla 1. Matriz de confusión para un sistema de clasificación binario. Fuente: Elaboración propia.

	Clase verdadera	
Clase estimada	Clase 0	Clase 1
Clase 0	TP	FP
Clase 1	FN	TN

- Verdadero positivo (TP): El número de muestras de la clase cero que el sistema clasifica exitosamente como pertenecientes a la clase cero.
- Falso negativo (FN): El número de muestras de la clase uno que el sistema clasifica exitosamente como pertenecientes a la clase uno.
- Falso positivo (FP): El número de muestras de la clase uno que el sistema clasifica erróneamente como pertenecientes a la clase cero.
- Falso negativo (FN): El número de muestras de la clase cero que el sistema clasifica erróneamente como pertenecientes a la clase uno.

Tasa de acierto (Acc): Es la medida de la cantidad de muestras que fueron clasificadas correctamente por el sistema, tanto de la clase cero como de la clase uno.

Sensibilidad (Sen): Mide la capacidad del sistema para clasificar adecuadamente las muestras de la clase cero o clase de referencia.

Especificidad (Spe): Mide la capacidad del sistema para clasificar adecuadamente las muestras de la clase uno. Dicho de otra forma, mide la capacidad del sistema para rechazar las muestras que no pertenecen a la clase de referencia.

Precisión: Mide la proporción de las muestras clasificadas como pertenecientes a la clase cero que verdaderamente pertenecen a ella

Recall: Mide cuántos de los elementos clasificados como de la clase de referencia efectivamente pertenecen a la clase cero. (Narkhede, S. (2018, May 9))

6. Bases de datos

Durante la investigación para este proyecto de grado, se hizo una búsqueda de bases de datos que fueron usadas para la solución de diversos problemas de

seguridad ciudadana en ciudades inteligentes. A continuación, se describen dichas bases de datos en compañía del problema que fue resuelto partiendo de ellas.

6.1. Datos abiertos de las autoridades de la ciudad de San Isidro, Lima, Perú

"Datos Abiertos" u "Open Data" (en inglés), es parte de una iniciativa mundial de Gobierno Abierto que busca que la información, en especial aquella que poseen las administraciones públicas, se publique de forma abierta, regular y reutilizable para que pueda ser empleada y redistribuida libremente por personas e instituciones, sin limitaciones ni licencias. (Portal de Datos Abiertos. 2016)

En este caso, los datos pertenecen a la sección "Seguridad y Control Ciudadano" la cual brinda datos sobre intervenciones policiales, apoyo operativo a la Policía Nacional del Perú, apoyo en auditorías, apoyo a bomberos, etc (G. B. Rocca, M. Castillo-Cara, R. A. Levano, J. V. Herrera and L. OrozcoBarbosa. 2016)

La aplicación utilizó la información de "ayuda a bomberos" para proporcionar a los peatones consejos útiles sobre la ruta más corta y segura a seguir para llegar a su destino final dada su ubicación actual. (G. B. Rocca, M. Castillo-Cara, R. A. Levano, J. V. Herrera and L. OrozcoBarbosa. 2016)

6.2. Catálogo de datos abiertos de la ciudad de Vancouver, Canadá

El catálogo de datos abiertos de la ciudad de Vancouver es un portal donde se almacenan 177 datasets de diferente índole entre las cuales se destacan salud, seguridad, arte, transporte, entre otros. (City of Vancouver Open Data Portal)

La base de datos de crímenes, la cual proporciona información sobre el tipo de delito cometido, la hora y el lugar del delito, además de una base de datos de vecindarios que contiene los límites de las 22 áreas locales de la ciudad en el sistema de información geográfico (SIG), fueron utilizados por Kim, Suhong (2018) para la generación de mapas de calor dentro de la ciudad de Vancouver.

6.3. Portal de datos abiertos de Queensland, Australia

El gobierno de Queensland se compromete a construir un ecosistema de datos confiable que haga que los datos importantes y no confidenciales estén abiertos para que cualquiera pueda acceder, usar y compartir. (Queensland Government. 2015)

El dataset de crímenes del portal de datos abiertos de la ciudad de Queensland proporciona información sobre la ubicación y el tiempo de ocurrencia de diferentes tipos de eventos delictivos. Hay un total de 17 tipos de eventos delictivos en este conjunto de datos. Rumi, S. K. (2018) utiliza este dataset para realizar predicción de crímenes.

6.4. Portal de datos abiertos de Chicago, Illinois

El portal de datos abiertos de la Ciudad de Chicago le permite encontrar datos de la ciudad, le permite encontrar datos sobre su vecindario, le permite crear mapas y gráficos sobre la ciudad y le permite descargar libremente los datos para su propio análisis. Muchos de estos conjuntos de datos se actualizan al menos una vez al día y muchos de ellos se actualizan varias veces al día (City of Chicago Open Data Portal. 2015)

El dataset de crímenes del portal de datos abiertos de la ciudad de Chicago contiene, entre otros, información relacionada con la hora, latitud/longitud del delito a nivel de manzana y uno de 27 tipos de crimen (ej, asalto o robo). Este conjunto de datos fue utilizado por Matthew S (2014) para predicción de crímenes y zonas de calor.

7. Metodología

Un sistema de clasificación binario es aquel en el se tiene la tarea de predecir variables categóricas, donde la salida está restringida a dos clases. Uno de los primeros pasos dentro del proceso es la extracción de los datos. En este trabajo se utilizaron dos bases de datos de la ciudad de Chicago, una de tweets y otra de

crímenes. Ambas bases de datos están estructuradas para el periodo de tiempo comprendido entre el 1 de enero de 2013 y el 31 de marzo del mismo año. La primera de ellas se extrajo desde la API que Twitter provee para investigadores y con la cual se pueden extraer tweets históricos y segmentar la consulta por espacio geográfico. El endpoint utilizado para la consulta fue https://api.twitter.com/2/tweets/search/all. Dicha consulta trajo como resultado 1.307.827 tweets, de los cuales se descartaron 278.774 porque no contaban con la información de latitud y longitud que son estrictamente necesarios para la implementación. En la **figura 1** se observa la distribución de los tweets mencionados (los tweets que se evidencian fuera de la ciudad de Chicago no son tenidos en cuenta).

La base de datos de crímenes se extrajo del portal oficial de datos de la ciudad de Chicago, donde se puede descargar la información de los crímenes cometidos desde el 2001. Cada registro contiene información como el tipo de crimen cometido, la fecha y la hora, la ubicación geográfica entre otros. Esta base de datos contiene 71.918 registros de los cuales se descartan 343 registros por no contener información geográfica. La distribución por cada tipo de incidente se muestra en la **tabla 2** y la distribución en el mapa en la **figura 2**.

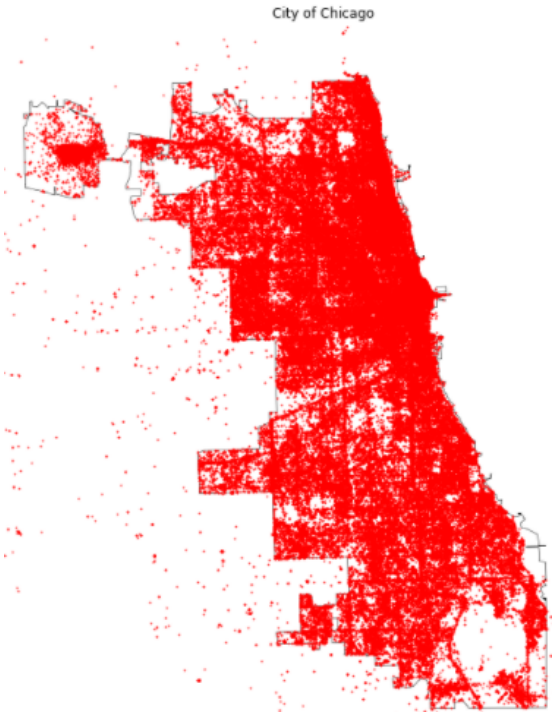

Figura 1. Distribución de los tweets realizados en la ciudad de Chicago entre enero y marzo del 2013. Fuente: Elaboración propia.

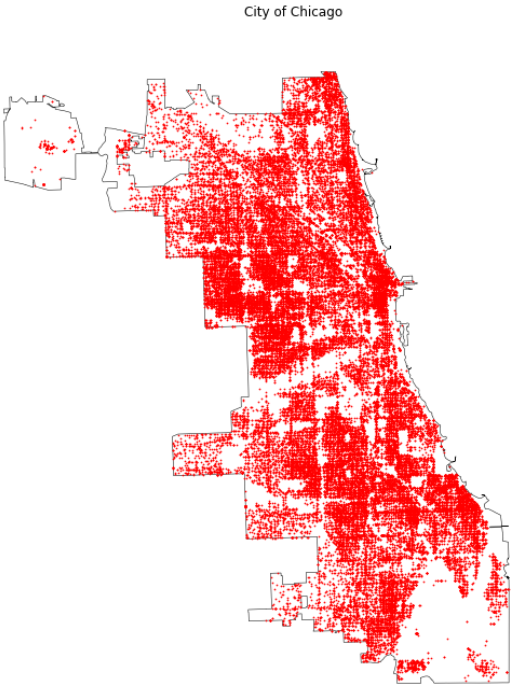

Figura 2. Distribución de crímenes en la ciudad de Chicago entre enero y marzo del 2013. Fuente: Elaboración propia.

El propósito de este trabajo es predecir los crímenes que van a suceder en un espacio y tiempo determinados, por lo tanto, fue imprescindible acotar el espacio geográfico sobre el cual se quería trabajar. No se tomó la ciudad completa porque, lo que sucede en el norte de ella, no necesariamente tiene correlación con lo que sucede en el sur. Así pues, se tomó la decisión de dividir la zona en áreas iguales de un kilómetro de ancho por un kilómetro de largo. Sobre ellas, se buscó cuál era el área en el cuál más crímenes se habían cometido durante los meses de interés. El resultado, como se esperaba, arrojó que la mayor cantidad de crímenes se cometieron en el centro de la ciudad como se muestra en la **figura 3** con un total de 878 crímenes (de todos los tipos).

Tabla 2. Distribución de los incidentes en la ciudad de Chicago entre los meses de enero y febrero del 2013. Fuente: Elaboración propia.

TIPO DE CRIMEN	CANTIDAD	TIPO DE CRIMEN	CANTIDAD
THEFT	15341	INTERFERENCE WITH PUBLIC OFFICER	295
BATTERY	11792	CRIM SEXUAL ASSAULT	293
NARCOTICS	9067	SEX OFFENSE	248
CRIMINAL DAMAGE	6873	LIQUOR LAW VIOLATION	138
OTHER OFFENSE	4721	ARSON	84
BURGLARY	4002	HOMICIDE	73
ASSAULT	3993	KIDNAPPING	64
MOTOR VEHICLE THEFT	3806	STALKING	32
DECEPTIVE PRACTICE	3378	GAMBLING	27
ROBBERY	2597	INTIMIDATION	26
CRIMINAL TRESPASS	2095	CRIMINAL SEXUAL ASSAULT	8
WEAPONS VIOLATION	781	OBSCENITY	5
PUBLIC PEACE VIOLATION	684	HUMAN TRAFFICKING	2
OFFENSE INVOLVING CHILDREN	664	NON-CRIMINAL	1
PROSTITUTION	485		

Una vez se escogió el área sobre la que se iba a realizar la predicción de crímenes, se procedió entonces a identificar la cantidad de tweets sobre la misma. Un total de 31.556 tweets se realizaron en el centro de la ciudad durante enero y marzo del 2013.

Teniendo definida el área, la cantidad de crímenes cometidos y de tweets realizados, se definió lo que serían las muestras que identifican un intervalo de tiempo específico. Una muestra está compuesta por los tweets realizados en un intervalo de tiempo de 500 minutos y está acompañada de una etiqueta que relata si en los 240 minutos posteriores hubo o no un crimen. Así, una ventana deslizante de 500 minutos se desplaza 240 minutos cada vez para generar todas las muestras que describen ambas bases de datos como se muestra en la **figura 4**.

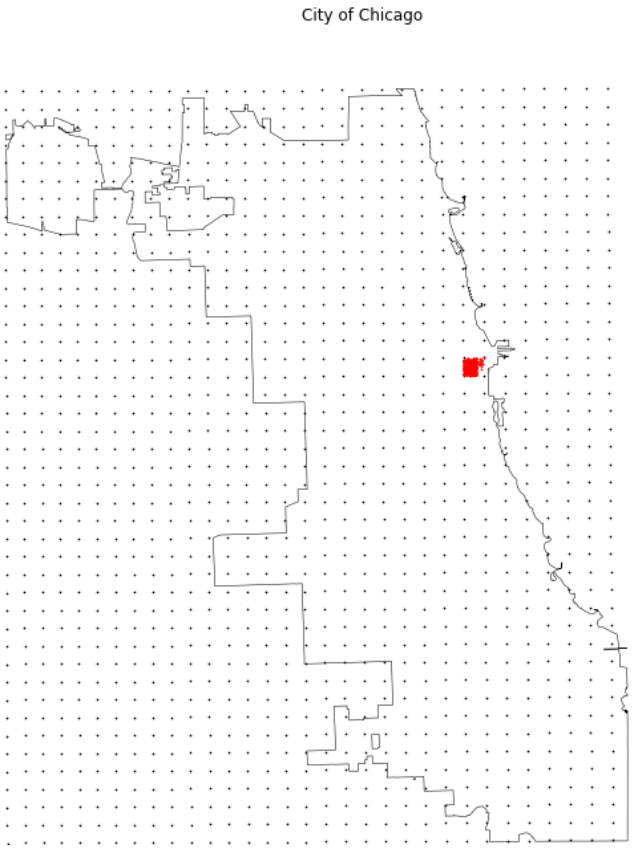

Figura 3. Delimitación de áreas y crímenes cometidos en el centro de la ciudad de Chicago. Fuente: Elaboración propia.

La decisión de los tiempos se realizó basándose en un análisis promedio de cada cuánto sucede un crimen en el centro de Chicago. Además, se buscó tener como resultado un dataset en el que la cantidad de muestras con crimen fueran suficientes para que el clasificador pudiera aprenderlas.

Se realizaron diversos experimentos donde se variaba el tiempo de cada muestra (tanto el tiempo para los tweets como el tiempo para los crímenes) y se encontró que la mejor combinación es la ventana propuesta.

Tomando este tiempo para la ventana de tiempo, se obtuvieron 537 muestras las cuales estaban distribuidas de la siguiente manera: 172 están clasificadas sin crimen y 365 están clasificadas con crimen.

Con las muestras expresadas en la forma anteriormente explicada no se puede entrenar un clasificador, por lo que se hizo necesario transformar los tweets en

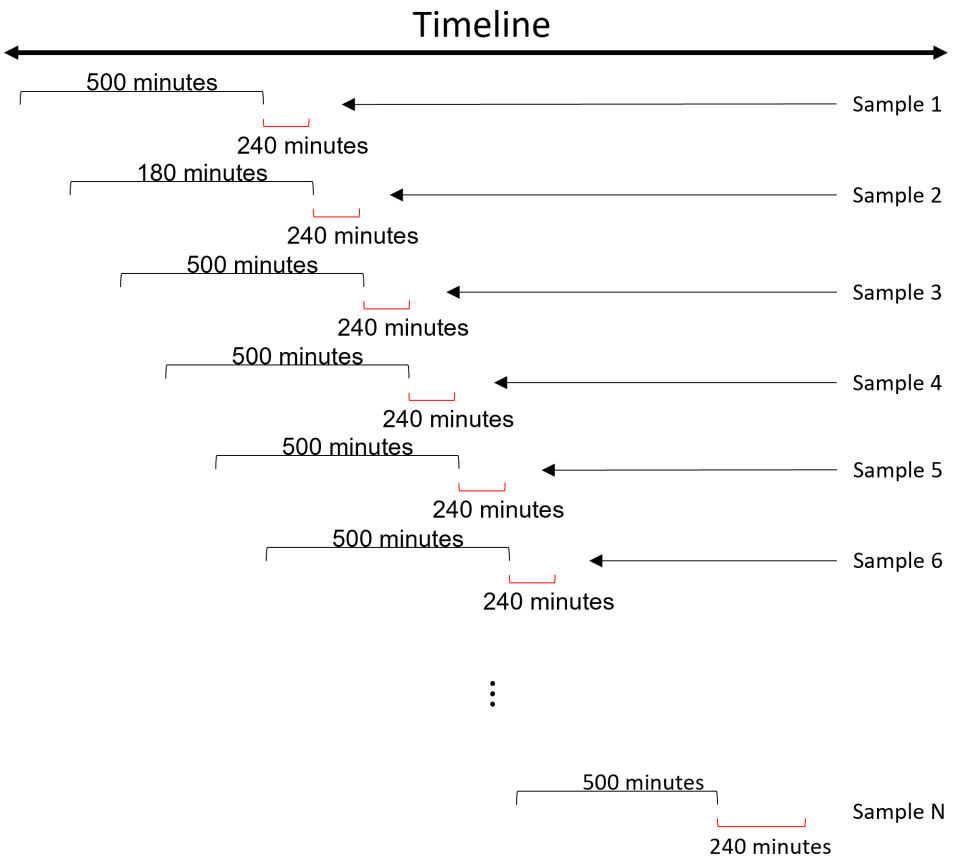

Figura 4. Extracción de las muestras a partir de las bases de datos de tweets y crímenes. Fuente: Elaboración propia.

características numéricas que los describan y que sirvan como conjunto de entrenamiento para el clasificador.

El primer paso para la extracción de características es la limpieza de los tweets. Twitter es una base de datos especialmente difícil lo que hace poco probable que un análisis semántico profundo de tweets a través de métodos tradicionales funcione bien (M. Gerber, J. Chai, A. Meyers. 2009).

Así pues, el primer reto fue realizar una limpieza adecuada para cada una de las muestras.

Aunque hay diferentes frameworks que ofrecen limpieza y tokenización para Twitter específicamente (A. Culotta, B. Huberman. 2010), se optó por hacer la limpieza manualmente para tener un mejor control sobre el proceso.

1. Se eliminaron todos los enlaces contenidos en las muestras
2. Se removieron todos los signos de puntuación, dentro de los que están contenidos los siguientes: !"#$%&\()*+,-./:;<=>?@[\\]^_`{|}~
3. Se hizo tokenización de las palabras
4. Se eliminaron las palabras vacías o 'stopwords'
5. Se aplicó lematización

En la **figura 5** se muestra una nube de palabras de los tweets antes y después de la limpieza y tokenización. Muchas de las palabras se conservan como se debería, pero se evidencia una cantidad menor de palabras vacías y también se puede observar el efecto de la lematización.

Figura 5. Nube de palabras del corpus previo a la limpieza (izquierda) y posterior a la limpieza (derecha). Fuente: Elaboración propia.

En la **tabla 3** se evidencian las cinco palabras más frecuentes encontradas en el corpus y en la **tabla 4** el proceso de limpieza y tokenización finalizado para una muestra.

Una vez que se tuvo cada una de las muestras limpia y tokenizada, era el momento de aplicar realizar la extracción de características.

Extracción de características:

Se extrajeron n características con el algoritmo Latent Dirichlet Allocation. En este punto era importante definir los hiper parámetros deterioro del aprendizaje (learning decay) y el número de tópicos.

Tabla 3. Visualización de tres tweets diferentes antes y después de la limpieza y tokenización. Fuente: Elaboración propia.

Palabra	Frecuencia
Chicago	15527
Job	11829
Il	7630
Get	5464
Ncaab	4410

Encontrar los valores adecuados para cada uno de los hiper parámetros que mejor se ajusten a los datos no es una tarea sencilla, por eso se tomó la decisión de implementar una búsqueda exhaustiva iterando sobre valores de parámetros específicos.

Luego de la búsqueda intensiva, se concluyó que los parámetros que permiten obtener el mejor rendimiento del modelo corresponden a $n = 10$ y 0.75 para el deterioro del aprendizaje.

Una vez el modelo estaba entrenado y se le entregaron las muestras tokenizadas, se visualizaron las díez primeras palabras para cada tópico para entender la distribución que el modelo hizo. Dicha visualización se muestra en la **figura 6**.

Tabla 4. Visualización de una muestra antes y después de la limpieza y tokenización. Fuente: Elaboración propia.

Muestra cruda	Muestra procesada
Case Management Coord II - #Chicago , IL (http://t.co/mFTT31FZ) Get Case Management Jobs #CaseManagement #jobs #job #GetAllJobs Certified Nursing Assistant - Bilingual - #Chicago , IL (http://t.co/kGqrqsWy) Get Psychiatry Jobs #Psychiatry #jobs #job #GetAllJobs The Roost Lodge - Where I stayed in Vail - 4-5-2012 for my first KINKY Concert ! I've been to a total of… http://t.co/zS2DoC3p Pharmaceutical S... - #Chicago , IL (http://t.co/j2mvTqwo) Get Pharmaceutical Sales Jobs #PharmaceuticalSales #jobs #job #GetAllJobs Housekeeping Assistant Manager - Ni... - #Chicago , IL (http://t.co/knpP343I) Get Healthcare Jobs #Healthcare #jobs #job #GetAllJobs Overland Contracting (Oci) - Le... - #Chicago , IL (http://t.co/nnWukNm5) Get Construction Jobs #Construction #jobs #job #GetAllJobs [NEWS] No. 2 Miami Hurricanes edge Virginia Cavaliers | http://t.co/xHmv5lva #ncaab #MIA [NEWS] Missouri Tigers rally, upend No. 4 Florida Gators | http://t.co/TDizkopY #ncaab #MIZZ [NEWS] No. 1 Indiana Hoosiers get win at No. 5 Michigan State | http://t.co/43uYARXZ #ncaab #MSU	['case', 'management', 'cooed', 'ii', 'chicago', 'il', 'get', 'case', 'management', 'job', 'casemanagement', 'job', 'job', 'getalljobs', 'certified', 'nursing', 'assistant', 'bilingual', 'chicago', 'il', 'get', 'psychiatry', 'job', 'psychiatry', 'job', 'job', 'getalljobs', 'roost', 'lodge', 'stayed', 'vail', 'first', 'kinky', 'concert', 'total', 'pharmaceutical', 'chicago', 'il', 'get', 'pharmaceutical', 'sale', 'job', 'pharmaceuticalsales', 'job', 'job', 'getalljobs', 'housekeeping', 'assistant', 'manager', 'ni', 'chicago', 'il', 'get', 'healthcare', 'job', 'healthcare', 'job', 'job', 'getalljobs', 'overland', 'contracting', 'oci', 'le', 'chicago', 'il', 'get', 'construction', 'job', 'construction', 'job', 'job', 'getalljobs', 'news', 'miami', 'hurricane', 'edge', 'virginia', 'cavalier', 'ncaab', 'mia', 'news', 'missouri', 'tiger', 'rally', 'upend', 'florida', 'gator', 'ncaab', 'mizz', 'news', 'indiana', 'hoosier', 'get', 'win', 'michigan', 'state', 'ncaab', 'msu']

```
Topic 0: de foto una art institute millennium wing gate cloud millenium
Topic 1: business pharmaceutical nd state library game big ul pharmaceuticalsales dj
Topic 2: state j k c r southern north florida saint texas
Topic 3: state league year bean happy contract j business english ice
Topic 4: management used tag consulting business managementconsulting supply accenture php supplychainmanagement
Topic 5: pitcher contract signed mlb bean nfl art millennium institute minor
Topic 6: management ecommerce business morning state deal assistant managementconsulting consulting con
Topic 7: nhl de forward business bean liga millennium ahl art go
Topic 8: league english united premier town two scottish bean super division
Topic 9: u class business cta morning go know got pic back
```

Figura 6. Primeras díez palabras para cada uno de los diez tópicos arrojados por LDA. Fuente: Elaboración propia.

Dando un repaso por cada uno de los tópicos, se puede evidenciar que, efectivamente, las palabras que en ellos se consolidan tienen una relación semántica. Por ejemplo, es fácilmente identificable que el primer tópico hace referencia a un lugar de la ciudad (millenium park) y actividades que en él se realizan.

Ahora bien, habiendo identificado los tópicos entre los cuales se distribuyen cada uno de los documentos (tweets que componen una muestra), se debe encontrar el porcentaje de pertenencia de cada documento a los tópicos. El resultado de esta operación entregará una matriz de tamaño NxM donde n es el número de muestras y m es igual a diez (número de tópicos).

Con las características extraídas anteriormente y cambiando las etiquetas categóricas a etiquetas binarias, donde 0 significa que no hubo un crimen en el intervalo de tiempo siguiente y 1 significa que sí lo hubo, se tiene una matriz de muestras/etiquetas de la forma en que se muestra en la **figura 7**.

$$
\begin{array}{ll}
\text{Sample 1} & \left[\begin{array}{cccccccc|c} f_{1,1} & f_{1,2} & f_{1,3} & f_{1,4} & f_{1,5} & f_{1,6} & \cdots & f_{1,M} & 1/0 \end{array}\right] \\
\text{Sample 2} & \left[\begin{array}{cccccccc|c} f_{2,1} & f_{2,2} & f_{2,3} & f_{2,4} & f_{2,5} & f_{2,6} & \cdots & f_{2,M} & 1/0 \end{array}\right] \\
\text{Sample 3} & \left[\begin{array}{cccccccc|c} f_{3,1} & f_{3,2} & f_{3,3} & f_{3,4} & f_{3,5} & f_{3,6} & \cdots & f_{3,M} & 1/0 \end{array}\right] \\
\text{Sample 4} & \left[\begin{array}{cccccccc|c} f_{4,1} & f_{4,2} & f_{4,3} & f_{4,4} & f_{4,5} & f_{4,6} & \cdots & f_{4,M} & 1/0 \end{array}\right] \\
\text{Sample 5} & \left[\begin{array}{cccccccc|c} f_{5,1} & f_{5,2} & f_{5,3} & f_{5,4} & f_{5,5} & f_{5,6} & \cdots & f_{5,M} & 1/0 \end{array}\right] \\
\text{Sample 6} & \left[\begin{array}{cccccccc|c} f_{6,1} & f_{6,2} & f_{6,3} & f_{6,4} & f_{6,5} & f_{6,6} & \cdots & f_{6,M} & 1/0 \end{array}\right] \\
 & \vdots \\
\text{Sample N} & \left[\begin{array}{cccccccc|c} f_{N,1} & f_{N,2} & f_{N,3} & f_{N,4} & f_{N,5} & f_{N,6} & \cdots & f_{N,M} & 1/0 \end{array}\right]
\end{array}
$$

Figura 7. Matriz de muestras/etiquetas extraídas de las bases de datos de tweets y crímenes. Fuente: Elaboración propia.

En este punto, ya se tenía todo lo necesario para comenzar a entrenar el modelo de clasificación. Antes de decidir cuál iba a ser ese modelo, se hizo un análisis estadístico de las características para decidir si todas ellas aportaban información importante o por el contrario se podrían descartar.

Primero, se hizo un análisis de correlación entre cada una de las once características mencionadas anteriormente. De dicho análisis se eliminaron las características que arrojaron un coeficiente de correlación superior a un número η. Este número η se escogió de un conjunto de valores con el cual se realizaron varias ejecuciones y se evidenció que el valor más adecuado era 0.7.

De las díez características propuestas, ninguna arrojó un coeficiente de correlación superior a 0.7, por lo tanto se entregan todas al clasificador.

En la **figura 8** se muestra el diagrama de cajas y bigotes para cada una de las características restantes

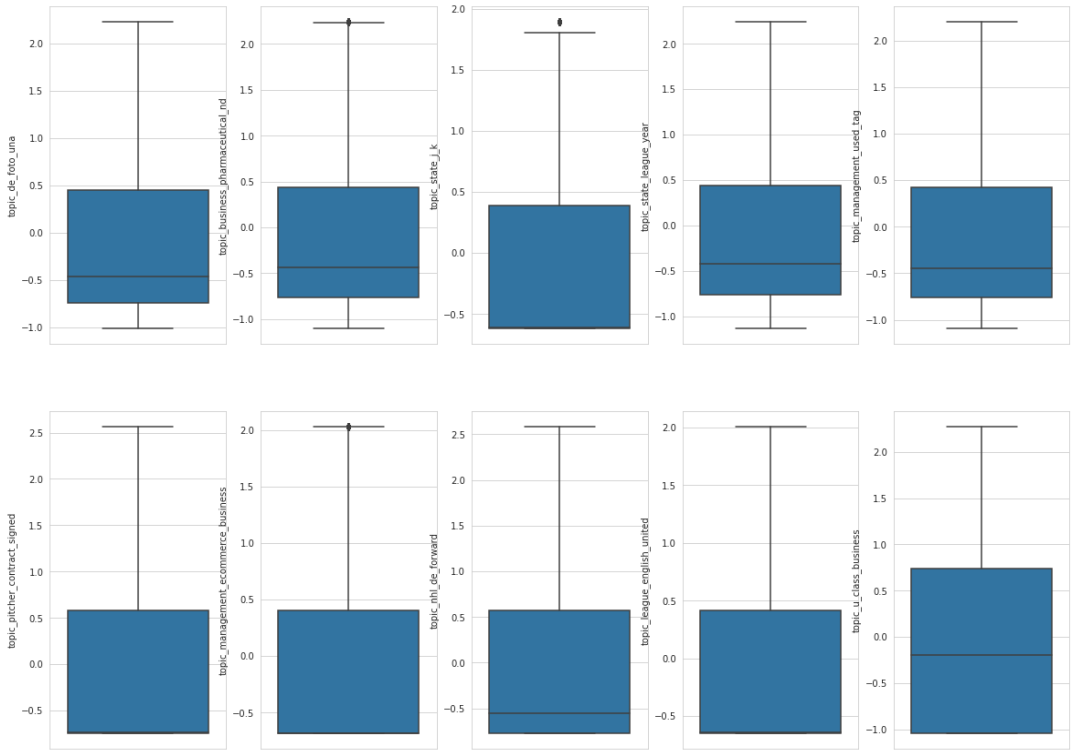

Figura 8. Diagrama de cajas y bigotes para las características seleccionadas. Fuente: Elaboración propia.

En el análisis del diagrama de cajas y bigotes arrojó que algunas características tenían algunos outliers dentro de su distribución. Dichos outliers fueron reemplazados por el valor máximo dentro de esa característica y con ello se limpió un poco más el conjunto de datos y se finalizó con el análisis estadístico de las características.

Una vez se tuvieron las características que serían presentadas al modelo, el conjunto de datos se separa en un 80% para entrenamiento y 20% para validación. El propósito de estos datos de validación es evaluar el comportamiento del modelo frente a datos no conocidos.

Para este caso se decidió utilizar una máquina de soporte vectorial como algoritmo de clasificación debido a que es un clasificador binario y se comporta muy bien en espacios de alta dimensionalidad. Además, la posibilidad de escoger entre diferentes kernel para hacer la transformación del conjunto de datos permite gran flexibilidad para tratar diversas distribuciones de datos. El hecho de poder controlar la compensación entre el tamaño del margen y la penalización de los puntos ubicados al otro lado del margen de decisión por medio del parámetro C, permite que se controle el sobre-entrenamiento. (C. M. Bishop. 2006)

El reto a partir de ese momento era encontrar los hiper-parámetros para la máquina de soporte vectorial que mejor se adaptaran a los datos y permitieran tener un mejor desempeño del modelo. Dichos parámetros se encontraron a través de una validación cruzada, donde se dividió el conjunto de entrenamiento en 10 grupos de entrenamiento y test con el propósito de que cada una de las muestras pasara por lo menos una vez por la fase de entrenamiento. Dentro del proceso de validación cruzada se entregó un conjunto de datos por cada hiper-parámetro que se deseaba tunear y el modelo arroja el conjunto de hiper-parámetros con el que se obtiene el mejor resultado para cada uno de los 10 grupos iniciales. Para los hiper-parámetros numéricos se obtiene el resultado final a través de la mediana y para los categóricos el que más se repite. En este caso, el valor para **gamma** es de 0.001098, **C** es 65.125180 y el **kernel** es sigmoide.

Finalmente, luego de tener un conjunto de parámetros definidos para la máquina de soporte vectorial, se evalúa el desempeño del modelo sobre los datos que nunca pasaron por la etapa de entrenamiento.

8. Resultados y análisis

Dentro de la metodología anterior, se entrenó una máquina de soporte vectorial con un conjunto de características extraídas de una base de datos de tweets y una base de datos de crímen. Estas características están basadas en un algoritmo de procesamiento del lenguaje natural que es Latent Dirichlet Allocation y una característica más producto de una hipótesis propuesta en este trabajo de grado.

En la **tabla 5**, se presentan los resultados obtenidos para la validación con el 20% de los datos que no fueron vistos por el clasificador. Estos resultados soportan la idea de que es posible predecir los eventos delictivos que van a ocurrir en un espacio y tiempo determinados a partir de los tweets que ocurrieron en el mismo espacio pero en un tiempo anterior.

La métrica más valorada en este caso es la especificidad (SPE), pues habla de la capacidad del sistema para decidir cuándo va a haber un evento delictivo. Dicha valoración se hace principalmente porque, en un caso de la vida real, que el sistema se equivoque y prediga que va a haber un crimen donde en realidad no lo va a haber, sólo causaría que las autoridades se alarmen, pero en cambio, si el sistema predice que no va a haber un crimen donde realmente sí lo habrá, podría traer graves consecuencias.

Tabla 5. Resultados de la clasificación. Fuente: Elaboración propia.

Característica	ACC (%)	F-Score (%)	SPE (%)	AUC
LDA	77	75	87	0.77

9. Conclusiones

En este trabajo se presentó una metodología para predecir crímenes en un espacio y tiempos determinados a partir de los tweets que se hicieron en ese mismo espacio y en un periodo de tiempo anterior. Es importante destacar el hecho de que es una herramienta poderosa para predecir y prevenir crímenes de diversos tipos. Para realizar este tipo de predicciones, es fundamental que en la ciudad en la cual se desea implementar la aplicación se popularice el uso de redes sociales como Twitter. El conjunto de datos con el que se entrenó el modelo se extrajo a partir de dos bases de datos de diferentes fuentes pero de la misma ciudad, una de ellas contenía los tweets realizados en la zona y la otra los crímenes. La extracción de las características se hizo a través de Latent Dirichlet Allocation y el algoritmo de clasificación fue una máquina de soporte vectorial.

Los resultados obtenidos apuntan a que, en efecto, es posible predecir los crímenes que se serán cometidos en ese tiempo y espacio seleccionados partiendo de las redes sociales como única fuente de información. Esto habla del poder que tienen las redes sociales para revelar información basada en el contenido que las personas allí publican y cómo esto se relaciona con eventos futuros.

Por último, aunque en el estado del arte haya aplicaciones que utilicen Twitter para realizar diferentes tipos de predicciones y una especialmente se enfoque en la 'predicción' de crímenes en la ciudad de Chicago, ninguna de ellas lo hace para periodos de tiempo tan cortos y tan precisos como los presentados en este trabajo.

Dentro de las posibilidades a futuro, está la implementación de nuevas características que acompañen las extraídas del procesamiento natural del lenguaje y permitan al modelo aprender nuevos patrones que caractericen de una forma más exacta los crímenes.

También estoy convencido de que el camino a seguir es disminuir el tiempo de la ventana para cada muestra y también reducir las dimensiones del espacio de interés. Disminuir ese tiempo permitiría a las autoridades ser más exactas y tener menos margen de error, pues un periodo de tiempo de dos horas es, bajo mi punto

de vista, aún bastante alto. El problema que esto acarrea es que la cantidad de crímenes no es tan alta como para hacer una segregación mayor y, de la misma manera, la cantidad de tweets en el espacio de interés tampoco lo permite.

Una aplicación más completa es la implementación de un modelo de clasificación por cada una de las zonas de interés. Cada uno de estos modelos estará desacoplado y será entrenado con los tweets generados únicamente dentro de su zona. El objetivo final sería una ciudad donde en cada una de sus zonas se realice predicción de crímenes.

10. Agradecimientos

Agradecimiento a Luis Alejandro Fletscher Bocanegra por guiarme y asesorarme en la realización de este trabajo, al componente de analítica por permitirme hacer parte del proyecto y a mis padres por acompañarme y apoyarme siempre.

11. Referencias Bibliográficas

Caragliu, A. & Del Bo, C. & Nijkamp, P., (2009) ."Smart cities in Europe," Serie Research Memoranda 0048, VU University Amsterdam, Faculty of Economics, Business Administration and Econometrics. https://ideas.repec.org/p/vua/wpaper/2009-48.html

S. Srivastava, A. Bisht and N. Narayan (2017), "Safety and security in smart cities using artificial intelligence — A review," 2017 7th International Conference on Cloud Computing, Data Science & Engineering - Confluence, 2017, pp. 130- 133, doi: 10.1109/CONFLUENCE.2017.7943136.

Latin American and Caribbean Institute for Economic and Social Planning (1997): Guía para la identificación, preparación y evaluación de proyectos de seguridad pública, LC/IP/L.149 (preliminary version) Santiago, Chile, July

Schank, R. C. (1991). Where's the AI?. AI Magazine, 12(4), 38. https://doi.org/10.1609/aimag.v12i4.917

A. Jain, Aishwarya and G. Garg. (2020), "Gun Detection with Model and Type Recognition using Haar Cascade Classifier," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), 2020, pp. 419-423, doi: 10.1109/ICSSIT48917.2020.9214211.

G. B. Rocca, M. Castillo-Cara, R. A. Levano, J. V. Herrera and L. OrozcoBarbosa. (2016), "Citizen security using machine learning algorithms through open data," 2016 8th IEEE Latin-American Conference on Communications (LATINCOM), 2016, pp. 1-6, doi: 10.1109/LATINCOM.2016.7811562.

Matthew S. Gerber, Predicting crime using Twitter and kernel density estimation (2014), Decision Support Systems, Volume 61, 2014, Pages 115-125, ISSN 0167-9236, https://doi.org/10.1016/j.dss.2014.02.003.

Rumi, S. K., Deng, K., & Salim, F. D. (2018). Crime event prediction with dynamic features. EPJ Data Science, 7(1), 43.

Helmstetter, S.; Paulheim, H. Collecting a Large Scale Dataset for Classifying Fake News Tweets Using Weak Supervision. Future Internet 2021, 13, 114. https://doi.org/10.3390/fi13050114

M. Gerber, J. Chai, A. Meyers. (2009), The role of implicit argumentation in nominal SRL, Proceedings of Human Language Technologies, The 2009 Annual Conference of the North American Chapter of the Association for Computational LinguisticsAssociation for Computational Linguistics, Boulder, Colorado, 2009, pp. 146–154.

A. Culotta, B. Huberman. (2010), Towards detecting influenza epidemics by analyzing Twitter messages, Proceedings of the First Workshop on Social Media Analytics, ACM, 2010, pp. 115–122.

By Great Learning Team -. (2020, October 16). Understanding Latent Dirichlet Allocation (LDA). Great Learning. Retrieved April 5, 2022, from https://www.mygreatlearning.com/blog/understanding-latent-dirichlet-allocation/

R, V. (2018, September 11). Feature selection — Correlation and P-value - Towards Data Science. Medium. Retrieved April 5, 2022, from https://towardsdatascience.com/feature-selection-correlation-and-p-value-da8921bfb3cf

C. M. Bishop (2006), Pattern recognition and machine learning. Springer.

Narkhede, S. (2018, May 9). Understanding Confusion Matrix. Medium. Retrieved April 5, 2022, from https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

Portal de Datos Abiertos. (2016). Municipalidad de San Isidro. Retrieved April 6, 2022, from http://msi.gob.pe/portal/innovacion/portal-de-datos-abiertos/

City of Vancouver Open Data Portal. Open Data Portal. https://opendata.vancouver.ca/explore/

Kim, Suhong & Joshi, Param & Kalsi, Parminder & Taheri, Pooya. (2018). Crime Analysis Through Machine Learning. 415-420. 10.1109/IEMCON.2018.8614828.

Queensland Government. (2015). Open Data Portal. https://www.data.qld.gov.au

City of Chicago Open Data Portal. (2015). Chicago Data Portal. https://data.cityofchicago.org