Análisis de redes sociales para mejorar la identificación de patrones de robo de vehículos

Alejandro Vásquez [*]
Rodrigo Joannon [*]
Richard Weber [**]

Resumen

Este artículo presenta una metodología para extraer, preprocesar y analizar datos de dos fuentes que captan denuncias de robos de vehículos. Una de las fuentes es la base de datos de la Asociación de Aseguradoras de Chile; la otra son los tuits que se publican en Twitter. Se muestra los sesgos que ambas bases tienen respecto del fenómeno de robo de autos y cómo los tuits podrán ayudar en la prevención de este delito. Dentro de los principales resultados encontrados, se identifica que existen diferencias en la proporción de denuncias dependiendo del valor del vehículo en las fuentes utilizadas. Sin embargo, para los modelos más robados, ambas fuentes presentan tasas de denuncia similares y, a su vez, los vehículos denunciados en la Twitter presentan mayor tasa de hallazgo que aquellos que no lo son independiente de su valor monetario. Este trabajo es uno de los fundamentos para la creación de un observatorio digital de apoyo a la industria automotriz chilena.

Palabras Clave: Patrones de crimen, Análisis de redes sociales, Robo de vehículos.

[*]Asociación de Aseguradores de Chile, Santiago, Chile
[**]Departamento de Ingeniería Industrial, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile, Santiago, Chile

1. Introducción

En Chile el robo de vehículos presenta tasas bastante elevadas, redondeando los 30.000 al año, lo que significa que un vehículo es sustraído cada 17 minutos. Este artículo pretende ayudar a entender este fenómeno y desarrollar un modelo predictivo del robo de vehículos utilizando dos fuentes de información: denuncias de vehículos y Twitter.

Twitter es una red social donde se comparten opiniones o declaraciones en tiempo real, los cuales son llamados "Tweets" [1]. Sus características han captado la atención de distintos investigadores que utilizan esta información para explicar fenómenos sociales [7].

El objetivo principal de este trabajo es identificar las relaciones existentes entre el robo de vehículos y denuncias realizadas a través de Twitter con el fin de establecer las diferencias y similitudes entre ambas fuentes de datos.

En general, se pudo mostrar que un análisis de los tuits aporta información al reconocimiento de patrones del fenómeno de robo de vehículos. Además se recreó las etapas del proceso por las cuales pasa la mayoría de los robos de vehículo y se pudo mostrar que la tasa de hallazgo es mayor en el caso de aquellos robos que fueron tuiteados.

El Capítulo 2 de este artículo describe tanto el problema de robo de vehículos como el uso de redes sociales para la prevención del crimen. En el Capítulo 3 se muestra cómo el análisis de tuits podrá ayudar a una mejor comprensión del fenómeno de robo de vehículos y presenta los resultados obtenidos. Finalmente el Capítulo 4 concluye el trabajo y señala trabajos futuros.

2. Definición del problema

El delito es una problemática social que se busca prevenir de la mejor forma posible. En Chile distintas instituciones gubernamentales están dedicadas a esta labor, como Carabineros de Chile, Policía de Investigaciones entre otros.

Existen más de 200 categorías de delitos, sin embargo se ha definido una categoría especial la cual incluye a las categorías de delitos que son más frecuentes en la sociedad, esta se llama "Delitos de Mayor Connotación Social" (DMCS) la cual incluye las siguientes categorías de delitos: Homicidio, Hurto, Lesiones, Violación, Robo con fuerza (Robo de accesorios de vehículos; Robo

[1] De aquí en adelante "tuits"; según la Real Academia Española.

de vehículo motorizado; Robo en lugar habitado; Otros Robos con fuerza), Robo con Violencia (Robo con intimidación; Robo con Violencia; Robo por sorpresa, Otros Robos con Violencia) [2].

Los DMCS son analizados y estudiados por el Departamento de Análisis Criminal de Carabineros de Chile (DAC). El 95 % de las denuncias son recibidas por Carabineros de Chile [20], el 5 % restantes son realizadas en Policía de Investigaciones (PDI) en su brigada especializada de robos.

Al momento de cuantificar la cantidad de delitos de alguna categoría existe el problema de la "cifra negra", definida como los delitos que son cometidos pero no denunciados. Para disminuir la incertidumbre de esta cifra negra Carabineros de Chile realiza encuestas periódicas en la población, además existe la Encuesta Nacional Urbana de Seguridad Ciudadana (ENUSC) [3], la cual tiene como objetivo obtener información sobre la percepción de inseguridad, la reacción frente al delito y la victimización de personas y hogares, a partir de una muestra representativa de zonas urbanas a nivel nacional y regional.

Según la ENUSC, el delito que menor tasa de cifra negra tiene es el "Robo o Hurto de vehículos" (3,6 %), por lo tanto es posible considerar que las cifras obtenidas de las denuncias para ese tipo de delito son representativas en cantidad y en fluctuaciones. La siguiente figura presenta el número de denuncias de robos de vehículos en un período de 5 años.

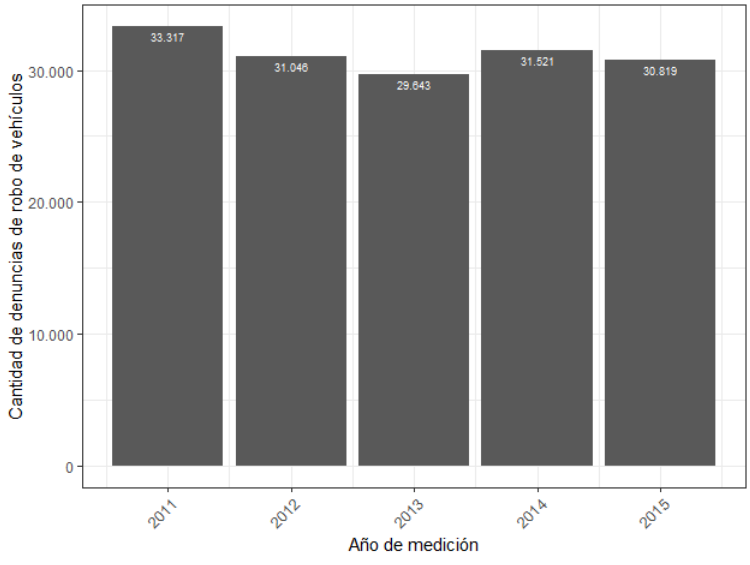

Figura 1: Denuncias de Robos de Vehículos. Elaboración propia con datos de "Estadísticas delitos de mayor connotación social [14].

El número de vehículos en circulación en Chile ha ido en constante creci-

miento, aumentando de 3.654.727 unidades en 2011 a 4.751.130 unidades en 2015 [5]. Cabe señalar que sólo el 30 % del parque automotriz cuenta con seguro particular, los cuales cubren principalmente daños y robo del vehículo. El robo de vehículos es el que involucra más costos para la aseguradora y es por eso que es de gran interés para estas compañías el poder evitar estos delitos.

2.1. Investigación del crimen usando Twitter

Twitter [19] es una plataforma de mensajes compartidos que es usado extensamente por las personas. Es una fuente de información muy enriquecida ya que los usuarios de la plataforma discuten públicamente hechos, emociones, y varios otros tópicos. La Figura 2 muestra un ejemplo de un tuit.

Figura 2: Ejemplo tuit. Fuente: Twitter [19].

El tuit considera los siguientes campos:

- Usuario: Nombre del usuario creador del tuit.
- Fecha: Fecha y hora en la cual fue elaborado el tuit.
- Contenido: El contenido expresado en el tuit contemplado dentro de los 140 caracteres máximos. Desde el año 2018 se permite 280 caracteres en un tuit. Dado que los experimentos descritos en este trabajo fueron ejecutados antes del año 2018 se usa a lo largo de este artículo 140 como número máximo de caracteres de un tuit.
- Acciones: Los tuits tienen acciones ejecutables por otros usuarios que visualizan dicho tuit. Estas acciones son: Reply, ReTweet y Like. La

primera contiene el número de respuestas que tuvo dicho tuit, el segundo es la cantidad de veces que fue compartido el tuit por otros usuarios y el tercero corresponde a la cantidad de usuarios que expresaron gustarle el contenido del tuit.

- Geolocalización: Ubicación GPS, desde dónde es emitido el tuit, y por ende es posible de rastrear desde qué ubicación geográfica el usuario escribió el mensaje. Sin embargo, este dato es opcional, los usuarios pueden optar por no entregar su posición GPS, y emitir el mensaje sólo con los otros campos.

Variados trabajos destinados a la investigación del crimen usando Twitter han mostrado que esta red social puede ser utilizada para describir o predecir comportamientos criminales en las ciudades. A continuación se revisará tres trabajos donde se propone usar Twitter para la identificación de fenómenos del crimen. Luego se mostrarán las conclusiones de este estudio bibliográfico para el trabajo presentado en el capítulo 3.

2.1.1. Revisión bibliográfica

En [11] se estudia el impacto del uso de redes sociales para estimar la tasa de crimen, en particular para hot spots dinámicos. En la investigación se analiza el riesgo de victimización criminal desde una dimensión espacial, es decir dependiendo de su localización. Para realizar el estudio sólo incluyeron mensajes en Twitter que tienen coordenadas GPS incorporadas.

Utilizaron dos cálculos de poblaciones para medir la población en riesgo de crimen, la población residencial y la población móvil. La que se utiliza más frecuentemente es la población residencial pero es muy poco probable que esta medida sea la adecuada para estimar la población en riesgo de crimen en aquellos delitos que involucran población móvil. Lo primero que realizan es utilizar la densidad de mensajes para marcarlos en un mapa y poder identificar rápidamente las zonas con más y menos cantidad de delitos. Luego comparan los resultados de ambas fuentes de datos.

Los resultados más llamativos son los de Leeds City Center, zona que tiene un gran volumen de eventos criminales violentos. Este no exhibe una tasa estadística significativa cuando se usa la población ambiente como la medida de población en riesgo. Una conclusión es que deben ser precavidos con el uso de los datos de Twitter y en hacer generalizaciones. Se sabe que algunos grupos socioeconómicos están sobre representados en los datos de Twitter. Adicionalmente, a pesar de que las tasas de usuarios de medios sociales están incrementando, el porcentaje de mensajes que incluyen información geográfica son cercanos al 1 %-2 %.

En [15] se usa Twitter para identificar hotspots del crimen. Realizaron una recopilación inicial de un mes de tuits, para comenzar el proyecto, y poder trabajar con datos de prueba, luego fueron ampliando la recopilación de tuits durante varios meses (enero a abril 2014). Utilizando los metadatos de localización que vienen vinculados con los tuits, distribuyeron geo-espacialmente en donde apuntaban esas referencias criminales para detectar donde hay puntos críticos de actos criminales. Para ubicar geo-espacialmente las referencias utilizaron la API de Google Maps para trazar ubicaciones. Por otro lado trazaron en un mapa los delitos reales, información obtenida por la Policía, ambos conjuntos de datos de la misma área de Inglaterra.

Realizaron ambos trazados con el fin de explorar la existencia de correlación entre ambas fuentes de datos. Se consideraron diferentes categorías de delitos, entre los cuales están: Comportamiento anti-social, robo de bicicletas, robo en propiedad, posesión de armas, orden público, hurto, robo a personas, robo de vehículos, crímenes violentos. Para cada tipo de crimen, identificaron palabras y frases claves relacionadas, las cuales eran relevantes para cada tipo de crimen en particular. Luego realizaron diferentes cálculos estadísticos.

En términos generales Twitter sub-estimaba considerablemente la cantidad de delitos, como por ejemplo en la categoría de "Comportamiento anti-social", la cual tenía una de las tasas más altas de denuncias. Si bien Twitter localizaba las mismas zonas con alta densidad, estas eran sub estimadas considerablemente, es decir en vez de abarcar una gran zona con alta densidad, detectaba varios puntos de esa zona. Otro ejemplo son los "Daños criminales", donde sucede lo mismo, y en donde los autores mencionan que en este caso esperaban mayor tasa de denuncias por Twitter ya que estos delitos son muy populares y están presentes en varias áreas. En el caso del robo de vehículos, prácticamente no detectaba zonas de alta densidad de denuncias, pero los autores explican que les fue muy difícil definir las palabras claves para identificar este tipo de crimen en particular y que puede ser que por haber escogido mal estas palabras es que no hayan detectado bien estas denuncias. Otro tema que menciona es la importancia de discriminar o categorizar de manera correcta los tuits, ya que influye directamente en los resultados. De hecho en el caso del robo de vehículos en particular, podría decirse que los resultados no son concluyentes ya que no pudieron captar una gran cantidad de tuits en esta categoría, y como mencionan los autores la justificación es que tuvieron dificultades para identificar cuáles tuits correspondían a esta categoría.

El artículo [4] muestra el uso de Twitter y "Kernel Density Estimation" (KDE) para predecir crimen. Recopilaron tuits entre enero y marzo del 2013, filtraron para aquellos que son emitidos dentro de la ciudad de Chicago y que tienen la opción de GPS activado. Consideraron 25 tipos de crímenes, entre los

cuales se considera robo, daños criminales, violación del uso de armas, asaltos, robo de propiedad privada, robo de vehículos, homicidios, entre otros.

El modelo utilizado para fue KDE, una técnica que ajusta un espacio bidimensional de función de densidad de probabilidad a un registro histórico de crímenes. Compararon dos tipos de modelos, el primero es el modelo que utiliza solamente características de KDE, el segundo es el modelo que combina KDE con los datos de Twitter.

Usaron una técnica de minería de textos para separar el tuit en palabras o tokens. Luego analizando esas palabras identificaron el tópico del tuit, asignando un tipo de crimen de las categorías opcionales. Para medir el rendimiento de los modelos, se grafica la capacidad de predicción del modelo, es decir se calcula AUC (area under curve) [1].

De los 25 tipos de crímenes considerados, 19 mostraron mejoras en la medición de AUC cuando se añadieron los tópicos de Twitter al modelo KDE. Dentro de los cuales está el robo de vehículos motorizados, el cual mostró una mejora de 0,69 a 0,71 en la medida AUC.

El modelo que cataloga los tuits en tópicos, es no supervisado, es decir que el humano no interviene en el aprendizaje del modelo, y según los autores esto dificulta el entender por qué algunos tipos de crímenes tienen mejores resultados que otros.

2.1.2. Conclusión de la investigación bibliográfica

Diferentes estudios e investigaciones han sido realizados en torno al tema central del crimen en las ciudades. Todas coinciden en la importancia que tiene categorizar bien los tuits en los tipos de delitos. Otra de las conclusiones más frecuentes en estas investigaciones es que es importante identificar los sesgos que presenta la plataforma social Twitter al ser utilizada como fuente de información de un modelo predictivo del crimen.

No pueden tampoco identificar cuánto de los errores cometidos en las predicciones pueden ser atribuidas a la red social por problemas de sobre exposición de ciertos perfiles de usuarios, ya que no hacen el análisis de los sesgos que presenta Twitter, pero mencionan que identificando estos sesgos y solucionando problemas mencionados en la investigación, Twitter sería una valiosa fuente de información para modelos predictivos.

Finalmente, al complementar una fuente de datos principal de denuncias reales de delitos con la fuente de datos compuesta por tuits, se concluye que los modelos predictivos mejoran su capacidad predictiva significativamente al incorporar a los tuits como una fuente de datos complementaria. Si bien al ser utilizada por si sola como fuente de datos principal no muestra la misma

capacidad predictiva que las denuncias reales efectuadas en los organismos responsables correspondientes, sí es una fuente de datos valiosa al ser considerada como complemento de otra fuente de datos principal, mejorando la capacidad predictiva de los modelos. Es así que dadas estas conclusiones, la motivación de este trabajo es avanzar en descubrir aquellas relaciones existentes entre una fuente de datos principal de denuncias realizadas formalmente y una fuente de datos secundaria basada en denuncias realizadas en Twitter con la finalidad de colaborar para el desarrollo futuro de modelos predictivos.

3. Usando Twitter para una mejor identificación del patrón de robo de vehículos

A continuación se mostrará la extracción de tuits. Luego se presentará los pasos del proceso Knowledge Discovery in Databases para preprocesar y analizar estos tuits.

3.1. Extracción de tuits

Para la extracción de tuits existen diferentes herramientas a utilizar, por un lado están las APIs ofrecidas por Twitter.

- AdsAPI (avisos): Destinada a la extracción de datos relacionados con los avisos publicitarios mostrados en Twitter [18].
- REST API (históricos): Permite extraer datos de Twitter históricos [16].
- Streaming API: Permite obtener tuits en tiempo real a través de fijar palabras filtros, todos los tuits escritos que contengan las palabras filtros serán recibidos por la API [17].

Por otro lado están los mecanismos de scraping: proceso que permite la extracción de una gran cantidad de datos de páginas web, el cual puede incluir únicamente el texto, todo el HTML de la página o incluso las imágenes dispuestas en la misma. Luego de un análisis de todas las opciones (ver [20]), se decidió usar la técnica de scraping para extraer los tuits relevantes en el periodo 2012 – 2016 donde se escogieron los siguientes términos de búsqueda:

robo patente OR robado patente OR robaron patente

La razón de haber escogido esos términos es porque en una primera etapa exploratoria se visualizaron diferentes tuits relacionados con robo de vehículos, y en todos se mencionaba la patente.

El objetivo de la extracción de tuits es obtener una base de datos de Twitter que represente denuncias de vehículos robados realizadas por la red social con el propósito de ser comparada con la base de datos de denuncias de AACH. Al utilizar los términos escogidos se podrá luego extraer de los tuits la patente del vehículo denunciado, la cual será un identificador único del vehículo denunciado por robo y a la vez será el dato clave para hacer la conexión con la base de datos de AACH. La patente no sólo es un dato importante por las ventajas que ofrece de poder obtener mayores datos con ella, sino que también es necesaria por las características de los tuits, los cuales presentan dos características que limitan la extracción de información correspondiente a las características de los vehículos.

La primera es la acotación del tuit a 140 caracteres, lo que provoca que las denuncias hagan referencias principalmente a características básicas del vehículo sin mayores detalles de cómo ocurre el acto delictual. Lo segundo es que al ser texto libre, muchas veces los términos usados para describir al vehículo son escritos de manera incorrecta o acotada. Sin embargo con la extracción de la patente denunciada es posible recuperar las características del vehículo, como el año, el modelo, el color o el tipo de vehículo.

3.2. Selección de datos

La base de datos de AACH consolida la información referente a las denuncias de robos de vehículos realizadas por las personas que sufrieron la sustracción de su vehículo entre los años 2012 y 2016 y que además cuentan con seguro particular (aproximadamente 30 % del parque automotriz).

La base original considera 45.018 registros con 74 atributos distintos, varios de los cuales remiten a la misma información sólo que usando nomenclaturas distintas por ejemplo comuna, y comuna id. De todos los atributos de la base, a priori se consideran los siguientes para la investigación:

- Aseguradora: Nombre de la aseguradora a la que está suscrita el vehículo sustraído, 13 categorías.
- Patente: Patente única del vehículo sustraído.
- Color: Color del vehículo, 16 categorías.
- Marca: Marca del vehículo, 107 categorías.
- Año: Año del vehículo.

- Estado: Busca o Encontró, indica si el vehículo fue encontrado o si aún está en búsqueda.
- Fecha último estado: Fecha del último estado, si el vehículo fue encontrado, este valor registra la fecha en que se encontró.
- Modelo vehículo: Registro del modelo, según el padrón del vehículo.
- Fecha denuncia: Fecha de cuando fue realizada la denuncia.
- Tipo de vehículo: Tipo de vehículo, 20 categorías. Separadas entre vehículos pesados, vehículos livianos u otros.
- Fecha Siniestro: Fecha y hora aproximada del robo del vehículo.
- Fecha AACH: Fecha de registro del robo en AACH.

Para realizar la extracción de tuits se aplicó la metodología descrita en el capítulo 3.1 considerando también el periodo entre el año 2012 y el año 2016 obteniendo 18.112 tuits.

La información de Twitter fue extraída de la red social como texto simple, por ende para obtener una base de datos con esa información es necesario distribuirla en variables como se muestra en el siguiente ejemplo de un tuit extraído.

> "Silvia PaillánCampos @kvyen 17 dic. 2016 Favor RT! robo de Mazfa Artis, 1999, color rojo burdeo, patente SU 3765. Si lo ven avisen al XXXXXXXXX o a carabineros al 133 RT! Responder Retwittear 163 Me gusta 9".

La minería de datos se aplica sobre el tuit, identificando caracteres especiales ocultos que segmentan la información almacenándola en variables. En el caso del ejemplo mostrado, las variables tendrían los siguientes datos:

> Usuario: "kvyen". Fecha: "17 dic. 2016". Hora: "7:37". (Este dato se obtiene de un dato oculto en formato time stamp). Texto tuit: "Favor RT! robo de Mazda Artis, 1999, color rojo burdeo, patente SU 3765. Si lo ven avisen al XXXXXXXXX o a carabineros al 133 RT!". Reply: "". ReTweet: "163". Like: "9".

Para realizar la extracción de las patentes contenidas en el texto de la variable "Texto tuit" se aplicó técnicas de minería de texto que se detalla en [20].

3.2.1. Limpieza y pre procesamiento

De los 45.018 registros de la AACH, 433 presentan registros irrecuperables, ya que no cuentan con la patente del vehículo ni modelo, por ende fueron eliminados, implicando que la base se redujera a 44.585 registros.

De los 44.585 datos, 1.744 registros presentaban datos faltantes como modelo del vehículo, tipo, año del vehículo o color. Para recuperar los datos faltantes se utiliza información obtenida por dos medios distintos: por la Ley N° 20.285 de Transparencia y por información de las plantas de revisión técnica.

Con Ley de Transparencia se entregó un listado de patentes y se obtuvieron datos de tipo de vehículo, año, marca y modelo. Con la información de Plantas de Revisión Técnica se obtuvieron los datos de color del vehículo. Entre los datos extraídos de Twitter que contienen la palabra "patente" se identificó 6.113 registros no válidos para este análisis ya que no tenían relación con robos de vehículos; por ejemplo:

> "Con la nueva patente de Apple en caso de robo el iPhone podrá recopilar información sobre el autor del mismo...".

Se aprecia que el tuit contenía las palabras filtros (robo + patente), sin embargo al aplicar el código de extracción de patentes hubiese extraído el término "de Apple" como posible patente de vehículo, la cual no es una patente válida, por lo tanto no corresponde a una denuncia de un robo de vehículo. Todos los tuits que fueron identificados con patentes no válidas fueron eliminados quedando 11.999 tuits válidos.

Paso importante en el pre procesamiento es la imputación de datos en el cual los valores faltantes ("missing values") o que han fallado alguna regla de edición del conjunto de datos son reemplazados por valores aceptables conocidos. La principal razón por la cual se realiza la imputación es para obtener un conjunto de datos completos y consistentes al cual se le pueda aplicar las técnicas de estadística clásica. Las razones para utilizar estos procedimientos en el análisis de datos son:

- Reducir el sesgo de las estimaciones.
- Facilitar procesos posteriores de análisis de datos.
- Facilitar la consistencia de los resultados entre distintos tipos de análisis.
- Mantener la estructura de asociación entre las variables.
- Mantener intervalos de confianza más robustos [12].

Al momento de imputar datos, los valores perdidos son llenados y la base de datos ya completada es analizada por métodos estandarizados. Los métodos comúnmente usados incluyen Hot Deck, Imputación por promedio e imputación por regresión [10].

Otro de los pasos fundamentales en el proceso de transformación, es la homologación de variables, es decir transformar la codificación de los datos de la variable de tal modo que los datos representen una categoría en particular que facilite su posterior análisis en el proceso de minería de datos. En ambas bases de datos se homologaron las variables con la misma codificación para que puedan ser relacionadas en el posterior análisis. A continuación se muestra algunos ejemplos de la homologación.

Color: Reducción a 1 palabra:

La variable color incorporaba distintas clasificaciones de colores, en más de la mitad de los registros estaba compuestos por 2 palabras por ejemplo "Gris Metalizado". Se decidió restringir el color a sólo 1 palabra, ya que ensuciaría el posterior análisis de esta variable. Gris Metalizado - Gris.

Modelos y Marcas: Codificación

Algunos modelos y marcas de vehículos presentaban 2 maneras distintas de referirse a una misma categoría por lo que se cambió su codificación por la categoría que presentaba mayor cantidad de registros. Por ejemplo: KIA - KIA MOTORS.

Tipo de vehículo: Codificación

Para la variable del tipo de vehículo fue necesario realizar distintos tipos de modificación en las categorías, principalmente agruparlas. Para definir las categorías principales se recurrió al archivo de tasación de vehículos del Servicio de Impuestos Internos (SII) [13].

Categoría AUTOMOVIL:

"AUTOMÓVIL" o "AMBULANCIA" o "COCHE MORTUORIO" o "TAXI BáSICO" o "LIMUSINA" o "JEEP"⇒ "AUTOMOVIL".

Categoría CAMIONES:

"TRACTO CAMION" o "TRACTOCAMION" o "CAMION" o "CAMIÓN" o "CHASIS CABINADO" o "CHASISCABINADO" o "CHASIS" ⇒ "CAMIONES".

La siguiente parte del proceso corresponde a la creación de nuevas variables que permiten realizar un mejor análisis de las relaciones existentes entre ambas bases de datos o mejorar las conclusiones posteriores.

Creación variable modelo:

Los modelos de los vehículos pueden presentar diferentes escrituras haciendo referencia al mismo vehículo, ya que para el proceso de inscripción no se sigue una nomenclatura. Además un modelo de vehículo puede presentar diferentes versiones ya sea con equipamiento "básico", o "top de línea". Otra complicación es que el modelo de vehículo puede presentar distintas versiones de potencia de motor, es así donde en los registros se encuentra por ejemplo:

CAMIONETA- NISSAN - 2010 - NAVARA 4X4 D/C DIES CUERO
CAMIONETA - NISSAN - 2010 - NAVARA 4X4 C/S DIESEL
CAMIONETA - NISSAN - 2010 - NAVARA 4X4
CAMIONETA - NISSAN - 2010 - NAVARA 4X4 D/C DIES AT CU

En donde los 4 vehículos son del mismo tipo "CAMIONETA", la misma marca "NISSAN", del mismo año, "2010", pero el modelo está escrito en 4 formas distintas ya sea por inscripción distinta o por ser distintas versiones. El principal problema de mantener estos 4 tipos de modelos es que dificulta el cálculo estadístico posterior, como tasas de frecuencias, ranking de modelos robados, o correlación con otras variables. Se toma la decisión de crear una variable restringida del modelo a sólo 2 palabras, de tal manera que los 4 ejemplos de modelos mencionados anteriormente quedarían registrados así:

CAMIONETA -NISSAN- 2010- NAVARA 4X4

Por lo tanto los 4 casos harían referencia a un mismo vehículo, si bien se pierden algunas características de la versión del vehículo, se asume el costo por el beneficio que se puede obtener en el análisis estadístico posterior.

Creación variable tasación:

Se decide crear una nueva variable con información de la tasación fiscal del vehículo, ya que esto permitiría ahondar en un mejor análisis respecto a segmentación de vehículos según tramos de valor de vehículo, y así encontrar los sesgos entre ambas bases de datos en cuanto a la valorización de los vehículos denunciados.

Para obtener la valorización fiscal de los vehículos se obtuvo la información histórica de valorización del Servicio de Impuestos Internos, el cual valoriza

anualmente a cada modelo de vehículo. La información se obtiene de manera online descargando los archivos del SII [13], los cuales contienen las siguientes variables: Tipo Vehículo – Marca – Modelo – Año – Valorización según año.

Creación variable grupo tasación:

Para evidenciar sesgos en cuanto a la tasación de los vehículos denunciados en ambas bases de datos era necesario obtener categorías comparativas o segmentos, es por esto que se decidió crear 4 segmentos o grupos de tasación. Considerando la base de datos de AACH como la base de datos de referencia, se calcularon los cuartiles de tasación, obteniendo:

1° cuartil (25 % de los registros): Valorización $\leq$ \$4.160.000

2° cuartil (25 % de los registros): \$4.160.000 $<$ Valorización $\leq$ \$5.480.000

3° cuartil (25 % de los registros): \$5.480.000 $<$ Valorización $\leq$ \$7.910.000

4° cuartil (25 % de los registros): \$7.910.000 $<$ Valorización

Con las reglas para obtener los cuartiles en AACH se creó una nueva variable que puede tener los valores 1, 2, 3 o 4. Según el cuartil al que corresponda. Esta variable fue creada tanto en la base de datos de AACH como en la base de datos de Twitter, pero utilizando en ambos casos el criterio recién mencionado, es decir el que define los cuartiles en la base de datos de AACH, considerando a esta base de datos como la base de datos de referencia.

3.2.2. Reducción de los datos

Según las características de las variables, los valores presentes en ellas, las dificultades posteriores que se podrían enfrentar si se mantiene la misma cantidad de datos, etc. se decidió eliminar determinados registros.

La primera reducción de datos realizada en las bases de datos fue eliminar los registros que no corresponden al tipo de vehículos "livianos", es decir todos aquellos registros en donde la variable "Tipo Vehículo" no contenga una de las siguientes opciones:

Automóvil - Camioneta - Furgón – Minibus - Moto - StationWagon – Todo Terreno.

Finalmente la decisión fue tomada considerando que los vehículos no livianos representan un 9 % de los datos tanto en la fuente de datos AACH como en la fuente de datos Twitter. Esta reducción de datos significó reducir de 44.585 registros a 41.692 en la base de datos de AACH y de 11.999 tuits a 11.025.

La base de datos de registros de denuncias de AACH presenta en este momento del proceso KDD 1.947 registros duplicados, los cuales fueron eliminados, manteniendo el que presentaba actualización más reciente.

En el caso de la base de datos de registros de denuncias realizadas por Twitter, se comprobó que también existen tuits que denuncian el mismo vehículo, esto se pudo comprobar a través de la variable patente, la cual fue extraída del tuit. Si bien el tuit puede no ser idéntico, están denunciando al mismo vehículo por lo que en este caso se mantuvo aquel tuit que fue emitido antes, es decir aquel que presenta mayor antigüedad. Según este procedimiento se identificaron 3.758 tuits que denuncian un vehículo ya denunciado anteriormente en otro tuit. La reducción de registros duplicados significó reducir de 41.692 registros a 39.745 (AACH) y de 11.025 tuits a 7.267.

En las tablas 1 y 2 se muestra un extracto de la base de datos de AACH y de Twitter (las patentes fueron modificadas).

3.2.3. Análisis estadístico

A continuación en el proceso KDD se aplicarán diferentes cálculos estadísticos que permitirán descubrir las relaciones entre ambas bases de datos, patrones y sesgos. Lo primero es realizar una inspección visual de las tasas de frecuencias históricas entre ambas bases de datos. Para confeccionar el gráfico de las tasas de frecuencias de denuncias, se acumularon mensualmente, con el fin de poder evidenciar con mayor claridad las relaciones existentes en cuanto al comportamiento de ambas fuentes de datos.

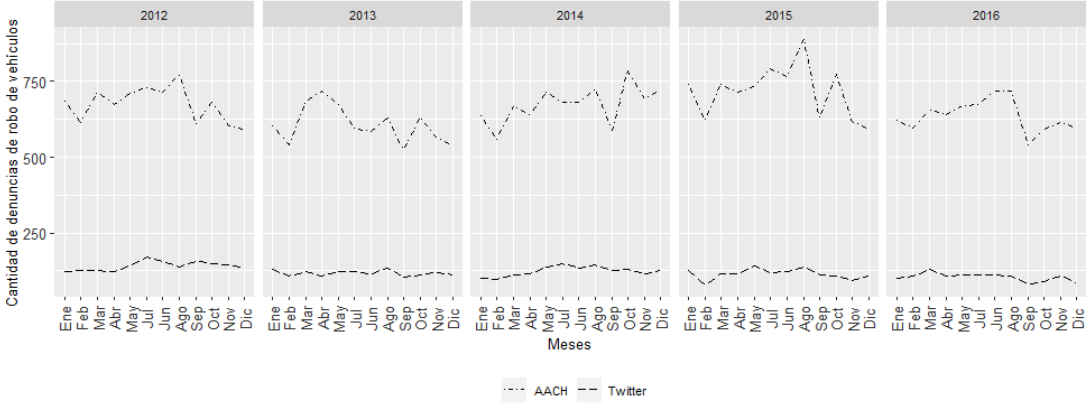

Figura 3: Gráfico Cantidad de denuncias 2012-2016.

Tabla 1: Base de datos AACH

Patente	Tipo Vehículo	Marca	Modelo	Año	Color	Tasación	Grupo	Fecha Sin	Fecha AACH	Fecha Hallaz	Si tweet
IJ1234	AUTOMOVIL	TOYOTA	YARIS GLI	2010	BLANCO	$5.090.000	2	27-01-2014	05-02-2014	29-01-2014	1
JK1234	STATION WAGON	KIA MOTORS	GRAND CARNIVAL	2012	BLANCO	$8.420.000	4	24-01-2014	03-02-2014	10-02-2014	0
KL1234	AUTOMOVIL	CHEVROLET	AVEO LT HB 5P1.4	2007	GRIS	$2.980.000	1	10-02-2014	19-02-2014	NA	0

Tabla 2: Base de datos Twitter

Patente	Tipo Vehículo	Marca	Modelo	Año	Color	Tasación	Grupo tasación	Reply	Retweet	Like	User	Fecha
AB1234	AUTOMOVIL	HONDA	CIVIC LSI 1.5	1995	VERDE	$1.250.000	1	1	41	6	Quinta_coquimbo	29-12-2016 9:39:36
BC1234	CAMIONETA	GREAT WALL	DEER 22	2008	BLANCO	$2.120.000	1	0	0	0	SilvPadilla	28-12-2016 19:13:12
CD1234	AUTOMOVIL	NISSAN	V16 SENTRA	2008	PLATEADO	$2.650.000	1	0	1	1	laarristaddet	27-12-2016 9:56:24

En Y están las cantidades de denuncias según la fuente de datos. En las frecuencias históricas se aprecia una cierta correlación en las tasas de frecuencias de ambas bases de datos, en particular en el periodo correspondiente a los años 2015-2016, en donde el crecimiento o decrecimiento de ambas bases de datos aparentan estar más correlacionados.

Un buen método para entender de manera numérica el grado de relación que tienen dos fuentes de datos distintas es aplicar un test de correlación. El coeficiente de correlación para el periodo 2012-2016 muestra 0,42 de correlación y 0,73 para el periodo 2015-2016.

Un segundo análisis está enfocado en identificar sesgos entre las bases de datos respecto de las características de los vehículos denunciados por robos. Para poder identificar los sesgos se realizan gráficos con las tasas de frecuencias de robos segmentado por grupo de tasación y modelo de vehículo. Las fuentes de datos consideradas son "AACH", Twitter, con todos sus registros, y por último "Twitter Aseg." que son los vehículos denunciados por Twitter que cuentan con seguros particular.

A continuación se presenta el gráfico de porcentaje de robos por **grupo de tasación** de vehículos, en donde se muestran las frecuencias que tienen cada una de las fuentes de datos.

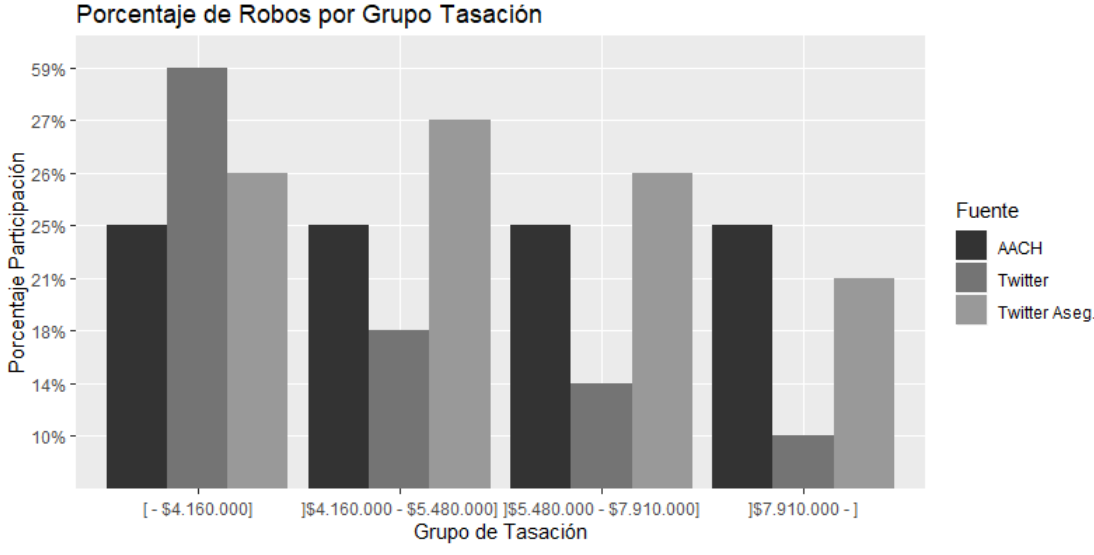

Figura 4: Gráfico Porcentaje robos por grupo tasación.

En el primer grupo de tasación, los de menor valor, Twitter presenta una altísima participación. Una de las justificaciones posibles, es que puede ser asociado a un segmento de menor poder adquisitivo.

La Figura 5 muestra el porcentaje de robos por **modelos de vehículos**,

en donde se presentan las frecuencias que tienen cada una de las fuentes de datos.

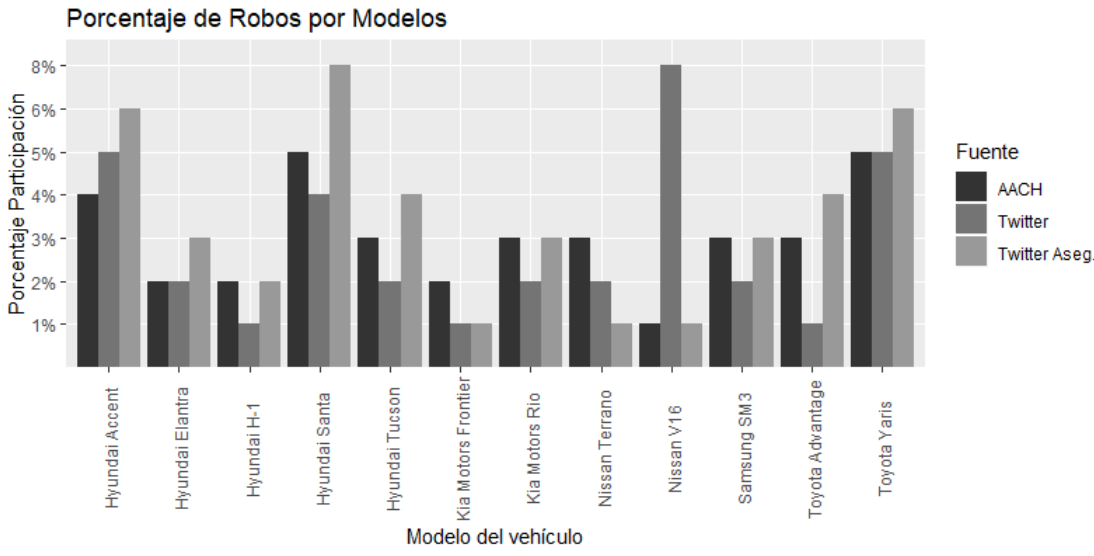

Figura 5: Gráfico Porcentaje de robos por modelos.

Una de las preguntas de investigación que estaba presente al momento de iniciar la investigación es de saber si el hecho de denunciar el robo de un vehículo por Twitter tiene un efecto positivo en la probabilidad de encontrar el vehículo.

Se verifica si el tamaño de muestra es suficiente para que sea representativo de acuerdo a la fórmula de Cochran [6, 8] y la distribución de los datos distribuye normal con el test prueba Kolmogorov-Smirnov [9].

Para esto se realiza un Test de Igualdad de Proporciones sobre las proporciones de vehículos robados que fueron encontrados, comparando entre aquellos que sólo fueron denunciados por las instituciones formales y aquellos que, además de hacerlo por esta vía, lo realizaron también por Twitter. Se define:

p_1: La proporción de vehículos robados que No fueron denunciados por Twitter y fueron encontrados.

p_2: La proporción de vehículos robados que Sí fueron denunciados por Twitter y fueron encontrados.

Hipótesis nula $H_0 : p_1 = p_2$
Hipótesis alternativa $H_1 : p_1 \leq p_2$

Como hipótesis alternativa se expresa que la proporción de hallazgos de aquellos vehículos robados que fueron denunciados por Twitter es mayor que aquellos que no lo fueron. Para analizar las tasas de hallazgo se consideraron en la base de datos de Twitter aquellos vehículos que están asegurados, es decir que las patentes coinciden en la base de datos de AACH, los cuales son 1.763 vehículos (24 % de los datos totales). Las tasas de hallazgos de ambas bases de datos son: 59 % (AACH) y 70 % (Twitter).

Al realizar un t-test de proporciones a un 99 % de confianza se obtiene como resultado un p-valor $2{,}2 \times 10^{-16}$. Algo importante de descartar es el hecho de que el sesgo en los grupos de vehículos denunciados en Twitter expliquen las diferencias en las proporciones, es decir que las tasas de hallazgos no mejoren en todos los grupos, sino que más bien se mantengan constantes pero que en Twitter se denuncien con mayor proporción aquellos vehículos que tienen mayor tasa de recuperación haciendo que la proporción general de hallazgo suba.

Por lo anterior es que se analiza las tasas de hallazgos entre grupos de tasación, para verificar si la proporción de hallazgos de los vehículos robados se ve afectada en todos los grupos. Para realizar esta comparación se consideran los registros de Twitter que también están presentes en AACH.

A continuación se presenta un resumen de los resultados obtenidos respecto al análisis de la frecuencia de robos y hallazgos de los vehículos robados según grupo de tasación y si fue denunciado en Twitter o no.

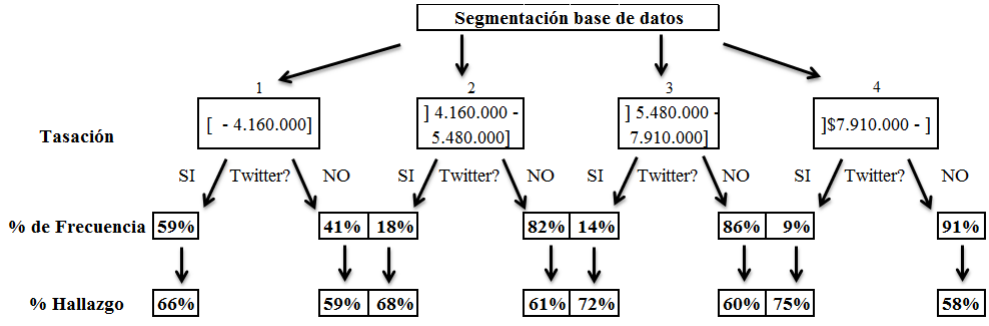

Figura 6: Resumen árbol de análisis según grupo de tasación.

La Figura 6 muestra cómo se generó una segmentación de las denuncias en 4 grupos de acuerdo a la tasación del vehículo, para luego seguir una segmentación según si el vehículo fue denunciado en Twitter o no, y según esta decisión medir la frecuencia con la que aparece este tipo de segmento en la fuente de datos y su correspondiente porcentaje de hallazgo. Recordar que el grupo de tasación se escogió dividiendo en 4 cuartiles la fuente de datos de

AACH de tal manera de que cada grupo representara un cuarto de la base de datos. Otra inquietud presente en esta investigación es saber si hay variación en las tasas de robos según el tipo de vehículo dependiendo el día de semana, lo cual hablaría de las intenciones detrás del robo. Para poder visualizar la variación se calculó la tasa de participación de robos de cada tipo de vehículo, y luego se calculó como esa tasa variaba dependiendo el día de la semana. Primero se analiza la fuente de datos de AACH y posteriormente la fuente de datos de Twitter.

En los resultados de AACH dos tipos de vehículos presentan variaciones importantes, el tipo de vehículo "Automóvil" y el tipo "Camioneta". La explicación podría residir en la intención del uso del vehículo robado, según Carabineros de Chile, algunos robos son realizados para ser utilizados para fiestas, en este caso privilegian los del tipo "Automóvil". Por otra parte algunos robos de vehículos se realizan con la intención de utilizar el vehículo para perpetrar un robo de otro tipo, como asaltos, o robos de bienes comerciales, para este caso es posible que los vehículos del tipo "Camioneta" sean más adecuados para cumplir la intención, además es común que las camionetas sean utilizadas con fines laborales y por ende con mayor frecuencia los días de la semana aumentando sus tasas de robos en estos días al estar más expuestas. Obtenido estos resultados provenientes de los datos de AACH, se mostrarán los resultados de Twitter para poder observar si estos muestran los mismos.

Los resultados de Twitter son similares, sólo que incorpora también el día viernes, pero el análisis sigue siendo el mismo, el fin de semana un tipo de vehículo "Automóvil" aumenta considerablemente y por el contrario el tipo de vehículo "Camioneta" decrece en los mismos días.

3.2.4. Interpretación de patrones

En este paso final se interpretan los patrones y relaciones encontrados en el proceso KDD. Se han aplicado diferentes técnicas que han permitido entender las relaciones entre ambas fuentes de datos. Previo a interpretar los resultados expuestos en el paso anterior se hará una reconstrucción del hecho, es decir, se ordenarán las fases por las que pasa un vehículo robado, concentrándose en vehículos asegurados.

- **Robo del Vehículo:** Corresponde al momento en que el vehículo fue robado. La hora de este evento es declarada por el dueño del vehículo al momento de denunciar en Carabineros de Chile, por lo que en algunos casos corresponde a una hora aproximada, ya que no siempre el dueño del vehículo puede observar cuándo el delito es realizado.

Tabla 3: Variación tasa de robo de vehículo según tipo, en fuente de datos Twitter.

Día	Automóvil	Station Wagon	Camioneta	Todo Terreno	Furgón	Minibus
lunes	-2,3 %	0,9 %	0,0 %	0,8 %	0,7 %	-0,2 %
martes	-0,1 %	-1,2 %	0,9 %	0,9 %	-0,3 %	-0,2 %
miércoles	-1,8 %	0,9 %	1,0 %	-0,2 %	0,2 %	-0,1 %
jueves	-0,7 %	-0,5 %	1,2 %	-0,1 %	-0,2 %	0,3 %
viernes	0,6 %	-0,3 %	-1,3 %	-0,3 %	0,7 %	0,5 %
sábado	4,3 %	-0,4 %	-1,7 %	-1,3 %	-0,6 %	-0,2 %
domingo	0,0 %	0,6 %	-0,2 %	0,2 %	-0,5 %	0,0 %
Promedio Anual	61,1 %	18,9 %	11,1 %	6,1 %	2,0 %	0,8 %

Tabla 4: Variación tasa de robo de vehículo según tipo, en fuente de datos AACH.

Día	Automóvil	Station Wagon	Camioneta	Todo Terreno	Furgón	Minibus
lunes	-1,7 %	-0,7 %	1,9 %	-0,3 %	0,9 %	0,0 %
martes	-2,0 %	-0,5 %	2,0 %	0,4 %	0,1 %	0,0 %
miércoles	-0,9 %	-0,7 %	1,5 %	0,0 %	0,0 %	0,1 %
jueves	-0,7 %	0,2 %	1,2 %	-0,3 %	-0,3 %	0,0 %
viernes	-0,8 %	0,6 %	0,5 %	-0,5 %	0,2 %	0,1 %
sábado	3,9 %	1,3 %	-4,6 %	-0,1 %	-0,4 %	0,0 %
domingo	2,7 %	-0,3 %	-3,0 %	1,0 %	-0,5 %	0,0 %
Promedio Anual	46,3 %	24,0 %	19,5 %	7,4 %	2,6 %	0,1 %

- **Envío del tuit:** Es la hora exacta en la que el tuit denunciando el robo del vehículo fue enviado.
- **Denuncia en Carabineros de Chile:** Es la hora en la que se realiza la denuncia formalmente en Carabineros de Chile por el robo del vehículo.
- **Validación AACH:** Es la fecha y hora en que AACH almacena los datos de las denuncias del robo del vehículo, la cual fue realizada en la aseguradora en la que el cliente está suscrito. Las aseguradoras no envían

los datos todos los días, sino que consolidan varias denuncias y luego se las transfieren a AACH, quien las valida y almacena.

- **Hallazgo del Vehículo:** Corresponde a la fecha y hora en la que el vehículo denunciado por robo fue hallado. Como no todos los vehículos son encontrados, para este cálculo se consideraron solamente aquellos que efectivamente fueron hallados.

Para medir los tiempos entre las fases se calculó la mediana de cada una. Asumiendo los tiempos medianos se puede reconstruir la secuencia de los hechos relacionados con un robo como lo muestra la Figura 7.

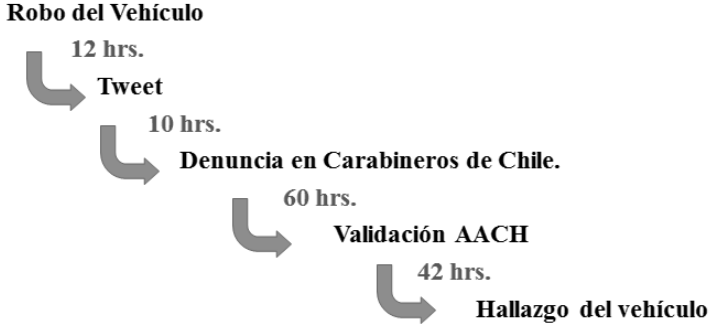

Figura 7: Reconstrucción orden de los hechos en un robo de vehículo.

Esto expresa el valor que presenta Twitter, ya que permite obtener información de manera anticipada, de hecho la mediana indica que el tuit se genera con 10 hrs. de anticipación frente a la denuncia formal. Además hay un largo periodo de tiempo transcurrido entre que el robo es realizado y AACH obtiene esa información, información que debe ser organizada y tratada para extraer los campos de interés. Además si bien en la generalidad de los casos, son aproximadamente 3 a 4 días los que transcurren desde que el vehículo es robado hasta que AACH obtiene la información, en muchos casos la obtención de esos datos se aproxima a las 2 semanas [20], por lo que en esos casos Twitter presenta un valor aun mayor al poder recibir información relacionada de manera anticipada. Respecto al tiempo transcurrido desde el robo hasta que es encontrado el vehículo, según la información de AACH, en la mayoría de los casos es cercana a los 5 días.

4. Conclusiones y Trabajo Futuro

Este trabajo presenta de obtención de tuits, extracción de información de ellos con técnicas de minería de texto, y aplicación de la metodología KDD sobre la fuente de AACH (datos de robos de vehículos asegurados) y sobre los datos de Twitter (tuits extraídos con un script).

Lo primero que se realizó en el proceso de minería de datos fue analizar la correlación entre las frecuencias de las denuncias de Twitter y AACH, el resultado mostró que tienen un comportamiento correlacionado entre el año 2012-2016, y que se acentúa más en el periodo 2015-2016 alcanzando un coeficiente de correlación de 0,73.

En cuanto a la valorización de los vehículos denunciados se expresó un gran sesgo, ya que en Twitter cerca del 60 % de los vehículos denunciados corresponden al grupo de menor valor, es decir aquellos que tienen una valorización menor a $4.160.000. Esto puede estar relacionado con el hecho de que quienes emplean mayor tiempo de navegación en internet son los más jóvenes, a quienes se les puede atribuir una menor disposición a pago por un vehículo. Lo importante de este hallazgo es considerar que al momento de analizar el robo de vehículos de alto valor, estos presentaran frecuencias considerablemente menores en Twitter, incluso pudiendo no ser factible de utilizar Twitter para predecir comportamientos delictuales para estos vehículos.

En cuanto a los modelos de vehículos denunciados, ambas fuentes de datos consideran tasas similares de robo.

Lo relevante del hallazgo es que para los modelos más robados, Twitter presenta tasas muy similares a las denuncias realizadas en AACH, y por lo tanto las estimaciones basadas en la red social deberían ser bastantes confiables para estos casos. Sin embargo, para los modelos de los vehículos utilizados para transporte de pasajeros no presentan correlación. Es fundamental ir monitoreando los modelos utilizados para estos fines ya que al momento de hacer las mediciones o utilizar el modelo del vehículo como variable predictiva se generarán conclusiones erróneas si no se considera.

Se descubrió que los vehículos denunciados en la red social presentan mayor tasa de hallazgo que aquellos que no lo son, en donde esta tasa de hallazgo no dependía de un sesgo asociado a la valorización del vehículo. La diferencia significativa de 11 puntos porcentuales puede indicar que el uso de las redes sociales ayuda a masificar las denuncias e incluso a recuperar sus pertenencias en caso de robo.

Este descubrimiento es altamente valioso para generar políticas públicas respecto al robo de vehículos, en donde se utilizan recursos de Carabineros de Chile para la recuperación de vehículos, ya que estimulando el uso de Twitter y otras redes sociales para denunciar los robos y posteriormente hallarlos se podría disminuir el costo en los recursos empleados actualmente para estos fines. Incluso las aseguradoras podrían estimular o promocionar el uso de esta red social para aumentar la probabilidad de encontrar un vehículo robado.

Se descubrió que el fin de semana el robo de vehículos de la categoría "Automóvil" aumenta y por otro lado disminuye aquellos de la categoría "Camioneta". Este patrón se evidenció en ambas fuentes de datos, lo cual es importante porque indicaría que Twitter no presenta sesgo en ese comportamiento. Finalmente se realizó una reconstrucción de las fases por las cuales pasa un vehículo robado en donde el resultado fue que primero el vehículo es robado, luego es denunciado por Twitter, para luego ser denunciado en Carabineros, siguiendo con la recepción de los datos por la institución AACH y finalmente el vehículo es hallado. Este ordenamiento mostró el valor que tiene Twitter como una fuente de información anticipada, ya que es el primer evento que sucede luego del robo del vehículo, de hecho es realizado antes que la denuncia formal en Carabineros de Chile. El hecho de que el tuit se origine como el primer evento luego de la sustracción del vehículo puede ser aprovechado por distintas entidades como AACH, municipios o las aseguradoras, para monitorear en todo momento los vehículos denunciados recientemente y aumentar las probabilidades de hallazgo al focalizarse en los que fueron denunciados en algún lugar en particular.

Tal como mostraban investigaciones del robo de vehículo que incluyen modelos predictivos, el no saber los sesgos en Twitter, dificultaba el entendimiento de algunos patrones de comportamientos, y dificultaba la optimización o mejoras en el rendimiento del modelo. Los descubrimientos aquí expuestos entregan información que permitirá estar conscientes de estos sesgos, pudiendo reparar errores en medición o en proporciones, por ejemplo en probabilidades de robos de algunos modelos.

En el trabajo futuro se construirá sobre la metodología desarrollada y los resultados presentados en este trabajo para construir un observatorio digital de robos de vehículos. El observatorio será alimentado con varias fuentes de información, entre ellas los siniestros guardados en las bases de datos de AACH y los datos publicados en redes sociales como por ejemplo Twitter.

Agradecimientos: Este trabajo fue financiado por el proyecto FONDEF ID16I10222 – CONICYT y el Instituto Sistemas Complejos de Ingeniería (ISCI) CONICYT FB0816. Se agradece el apoyo de la Asociación de Aseguradores de Chile (AACH) y de Carabineros de Chile.

Referencias

[1] B. Baesens. *Analytics in a big data world: The essential guide to data science and its applications*. John Wiley & Sons, 2014.

[2] Carabineros de Chile. Algunas definiciones : Delitos de mayor connotación social. `http://dac.carabineros.cl/datos.php`. En línea; Último acceso Diciembre 2016.

[3] ENUSC. Presentación de resultado xii encuesta nacional urbana de seguridad ciudadana, 2015. [Último acceso: Diciembre 2016].

[4] M. S. Gerber. Predicting crime using twitter and kernel density estimation. *Decision Support Systems*, 61:115–125, 2014.

[5] INE. Ine anuarios parque de vehículos en circulación 2015. `"http://www.ine.cl/canales/chile_estadistico/estadisticas_economicas/transporte_y_comunicaciones/parquevehiculos.php#"`, apr 2017. [En línea; Último acceso Agosto 2018].

[6] G. Israel. Determining sample size, arlington: Program evaluation and organizational development, ifas, university of florida. peod-6. *National Science Foundation, Research and Development in Insutry*, 1992.

[7] J. Kalyanam, M. Quezada, B. Poblete, y G. Lanckriet. Prediction and characterization of high-activity events in social media triggered by real-world news. *PloS one*, 11(12):e0166694, 2016.

[8] E. Lehmann y J. Romano. *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

[9] H. W. Lilliefors. On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American statistical Association*, 62(318):399–402, 1967.

[10] R. J. Little y D. B. Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2002.

[11] N. Malleson y M. A. Andresen. The impact of using social media data in crime rate calculations: shifting hot spots and changing spatial patterns. *Cartography and Geographic Information Science*, 42(2):112–121, 2015.

[12] D. Padula y L. Debera. Técnicas de imputación, una aplicación para medir el ingreso. In *Décimo Congreso Latinoamericano de Sociedades de Estadística. Córdoba*, 2012.

[13] Servicio de Impuestos Internos. Sii tasación fiscal de vehículos. "http://www.sii.cl/pagina/actualizada/noticias/tasacion_vehiculos.htm". [En línea; Último acceso: Abril 2017].

[14] Subsecretaría de Prevención del Delito. Delitos de mayor connotación social - series de datos 2001 - 2016, apr 2017.

[15] R. Todd-Bennet. Identifying crime hotspots using twitter: School of Computer Science and Informatics Cardiff University, 2015. [En línea; Último acceso Agosto 2018].

[16] Twitter. REST API. "https://developer.twitter.com/en/docs". [En línea; Último acceso: Agosto 2018].

[17] Twitter. Streaming API's. "https://developer.twitter.com/en/docs/accounts-and-users/subscribe-account-activity/guides/account-activity-data-objects1". [En línea; Último acceso Agosto 2018].

[18] Twitter. Twitter ads API. "https://developer.twitter.com/en/docs/ads/general/overview". [En línea; Último acceso: Agosto 2018].

[19] Twitter. Blog twitter. "https://twitter.com/twittersupport/status/555076845293432834", jan 2015. [En línea; Último acceso: Agosto 2018].

[20] A. Vásquez. Análisis de relaciones existentes entre datos de robos de vehículos e información extraída de twitter aplicando kdd. Memoria para optar al título de ingeniero industrial, Universidad de Chile, 2017.