UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA CIVIL INDUSTRIAL

ANÁLISIS DE RELACIONES EXISTENTES ENTRE DATOS DE ROBOS DE VEHÍCULOS E INFORMACIÓN EXTRAÍDA DE TWITTER APLICANDO KDD

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

ALEJANDRO ANDRÉS VÁSQUEZ CÁCERES

PROFESOR GUÍA:

RICHARD WEBER HAAS

MIEMBROS DE LA COMISIÓN:

IGNACIO CALISTO LEIVA

ROCÍO RUIZ MORENO

SANTIAGO DE CHILE

2017

RESUMEN DE LA MEMORIA
PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL INDUSTRIAL
POR: ALEJANDRO ANDRÉS VÁSQUEZ CÁCERES
FECHA: 14-08-2017
PROF. GUÍA: SR. RICHARD WEBER

ANÁLISIS DE RELACIONES EXISTENTES ENTRE DATOS DE ROBOS DE VEHÍCULOS E INFORMACIÓN EXTRAÍDA DE TWITTER APLICANDO KDD

Los delitos en cualquier sociedad son una problemática que se intenta disminuir lo mayor posible. 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 en Chile.

Esta investigación se enmarca en el proyecto de análisis de robos de vehículos desarrollado por PROSE, consultora que ofrece servicios de análisis y desarrollo a las empresas de seguros de vehículos de Chile, en conjunto con la Universidad de Chile y con el apoyo del Fondo de Fomento al Desarrollo Científico y Tecnológico.

El proyecto pretende ayudar a entender y desarrollar un modelo predictivo del robo de vehículos utilizando 3 fuentes de información, denuncias de vehículos, medios noticiosos y Twitter.

Twitter es una red social donde se comparten opiniones o declaraciones en tiempo real, los cuales son llamados Tweets, sus características han captado la atención de distintos investigadores que deciden utilizar esta información para explicar fenómenos sociales.

Se han presentado problemas de interpretación de patrones o rendimiento de modelamientos predictivos al utilizar Twitter debido a sesgos inmersos en la red social, producidos por las características de los usuarios, como de localización, edad, contenido compartido o intereses.

Es por esto que el objetivo principal de esta tesis es identificar las relaciones existentes entre el robo de vehículos y denuncias realizadas a través de Twitter en el periodo 2012 al 2016 con el fin de establecer las diferencias y similitudes entre ambas fuentes de datos.

Para el desarrollo de esta tesis se consideran dos fuentes de datos, las denuncias de robos de vehículos almacenados por PROSE y denuncias realizadas por Twitter en el periodo 2012 al 2016.

Se analizó la correlación de las denuncias entre ambas fuentes de datos, identificando que tienen una correlación de 0,73.

Se recreó el proceso por los cuales pasa un robo de vehículo, en donde se descubrió que el orden es 1° Robo del vehículo, 2°Envío del Tweet, 3°.Denuncia en Carabineros de Chile, 4°Validación PROSE, y finalmente el Hallazgo. Es decir el Tweet es el primer evento que se origina luego del robo. Además se descubrió que aquellos vehículos que son denunciados por twitter presentan tasas de hallazgo superiores, conocimiento valioso para futuros trabajos.

Algunos de los sesgos encontrados fueron que en Twitter se denuncia en una mayor proporción los vehículos de menor valor, además de mostrar que los modelos denunciados tienen tasas de frecuencias similares en ambas fuentes de información.

Agradecimientos

Con esta tesis se acaba un largo periodo en la universidad, el cual se llenó de desafíos desde el primer día de clases.

En los primeros años tuve complicaciones con ciertas materias que generaban dudas sobre si haber escogido esta carrera era la mejor decisión, sin embargo conté con el apoyo de diferentes personas que estimularon la decisión de seguir adelante y que hoy agradezco enormemente porque me doy cuenta que fue la mejor opción.

Estuvieron también aquellos que hicieron del proceso universitario una gran experiencia, los amigos y grupos de trabajo, con los cuales logré grandes aprendizajes y momentos de gran felicidad.

Quiero agradecer a mi familia en general por siempre mostrar interés en cómo estaba yendo mi carrera universitaria, y todos los consejos que me entregaron.

Agradezco especialmente a mis padres, quienes han forjado las raíces de lo que hoy soy, quienes me apoyaron en todos los procesos, por ser mis amigos en los momentos de mayor amargura. Este proceso estuvo muy lejos de ser fácil, sin embargo hoy se termina y son ellos otros responsables de que esto se haya cumplido.

Agradecer también a mis hermanas, por ayudarme siempre en lo que han podido, y hacerme saber que cuento con ellas en todo momento.

Quiero agradecer a todas las amistades que generé en la Universidad, ellos hicieron todo más fácil, hicieron que hasta en los momentos más complicados uno pudiera sacar una sonrisa. En especial quiero agradecer a Esteban Castro, Felipe Cortines, Fernando Brierley y Víctor San Martín, quienes fueron mis amigos más cercanos.

Quiero agradecer a Texia, por acompañarme, siendo siempre una motivación extra durante el final del proceso.

Y finalmente al profesor Richard Weber por haber guiado este trabajo, por contar con su apoyo incondicional, y por estar siempre atento al desarrollo de este.

Tabla de Contenido

Lista de Figuras

Lista de Tablas

Capítulo 1: Introducción

1.1 Proyecto Observatorio del fenómeno de Robo de vehículos en Chile.

PROSE es una consultora que ofrece servicios de análisis y desarrollo a las empresas de seguros de vehículos de Chile.

Hoy en día la principal problemática que enfrentan las aseguradoras es el robo de vehículos, por el alto costo que esto implica. Es por esto que PROSE en su interés por entender mejor el fenómeno del robo de vehículos desarrolla un proyecto llamado "Observatorio de robo de vehículos" en conjunto con la Universidad de Chile y con el apoyo del Fondo de Fomento al Desarrollo Científico y Tecnológico.

El observatorio recopila información de 3 fuentes distintas de información:

1. Datos de denuncias de robos de vehículos obtenidas a través de la información que brindan las aseguradoras a PROSE.
2. Datos de noticias online o páginas de internet con información de robos de vehículos.
3. Datos extraídos de Twitter.

El objetivo del proyecto es poder entender mejor y de forma anticipada el desarrollo del robo de vehículos en Chile para poder evitar su masificación a través del análisis desarrollado a las fuentes de información recopilada por el observatorio.

El modelamiento del proyecto es el siguiente:

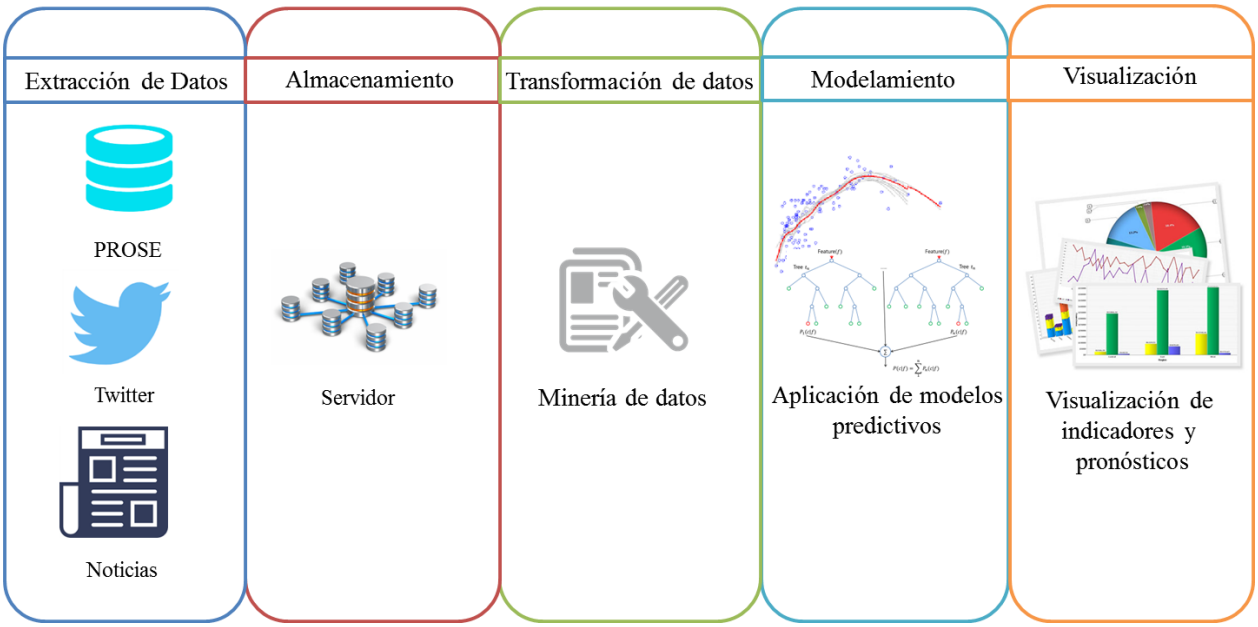

Figura 1.1: Funcionamiento Observatorio.

La primera fase es la extracción de los datos, la segunda fase es el almacenamiento, la tercera es la transformación de los datos almacenados para facilitar la aplicación de estadística sobre ellos, la cuarta fase es el modelamiento predictivo, el cual se aplica sobre los datos transformados, además de aplicar diferentes medidas estadísticas, y por último es la fase de visualización en donde se observan patrones de comportamiento y diferentes indicadores.

El proyecto tiene una primera duración de 2 años de desarrollo en donde se deben entregar los primeros avances que demuestren que con 2 años más de desarrollo será posible obtener un observatorio con modelamiento predictivo.

En esta primera etapa de desarrollo se deben entregar antecedentes de investigación que validen el modelo, análisis cualitativos y cuantitativos de las bases de datos a utilizar y maquetas de visualización del entregable final.

Cada una de las fases del proyecto descritas en Figura 1.1 se trabajan en paralelo hoy en día en equipos de trabajo guiados por profesores de la universidad de los departamentos de Ingeniería Civil Industrial y del departamento de Ingeniería Civil en Computación.

Esta investigación se focaliza en la fase de modelamiento, ya que los resultados serán utilizados para apoyar el modelamiento predictivo.

1.2 Chile y las redes sociales.

En Chile el número de acceso a internet (fijo + móvil 3G y 4G) ha superado los 13 millones, lo que se traduce en que la penetración de uso de internet ha alcanzado 72,4 accesos cada 100 habitantes [1].

La distribución de tiempo de navegación en internet por edad según datos del año 2014 es la siguiente:

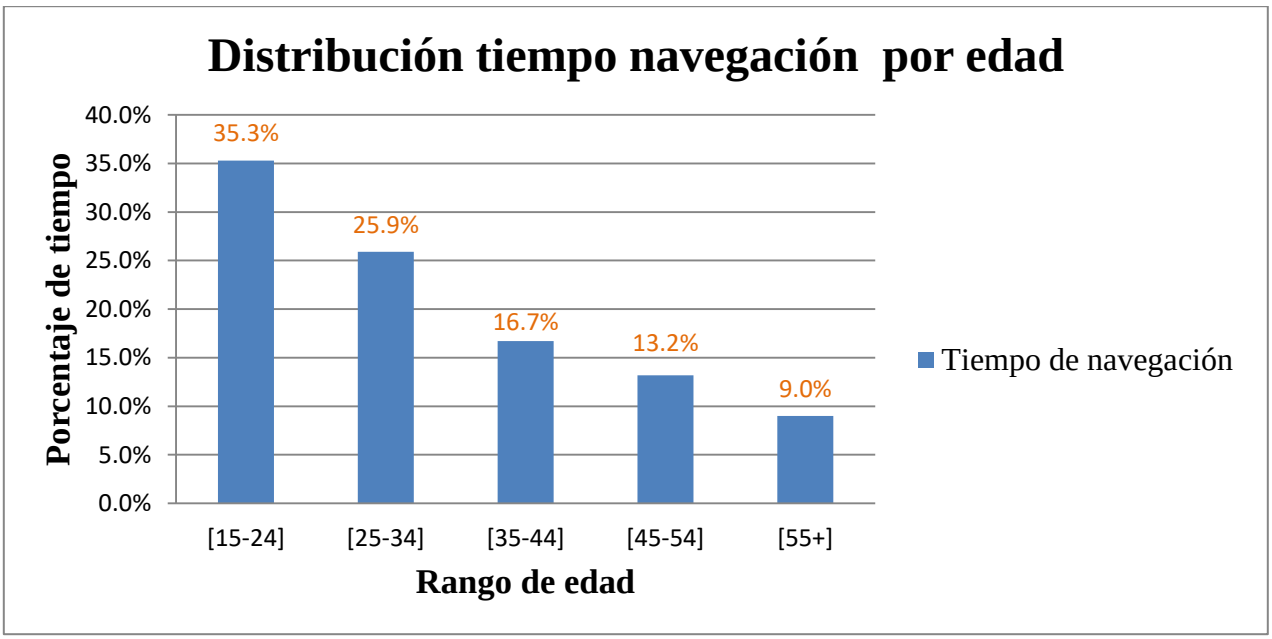

Figura 1.2: Distribución tiempo navegación por edad.
Fuente: comScore, Futuro Digital Chile [2].

Basado en la información del gráfico, el 61% del tiempo de navegación en Internet es empleado por usuarios cuya edad se encuentra en el rango de 15-34 años y el 79% se encuentra en el rango de edad de 15-44 años.

En cuanto a la navegación de internet, hoy en día a lo que más destinan tiempo los chilenos que ingresan por un dispositivo celular es a las redes sociales, superando incluso al navegador de internet [3].

Tal vez el hecho de que las redes sociales sean tan visitadas tiene relación con la edad de los navegadores preferenciales que tienen entre [15-24] años.

El siguiente gráfico presenta datos según usuarios únicos conectados por dispositivos móviles:

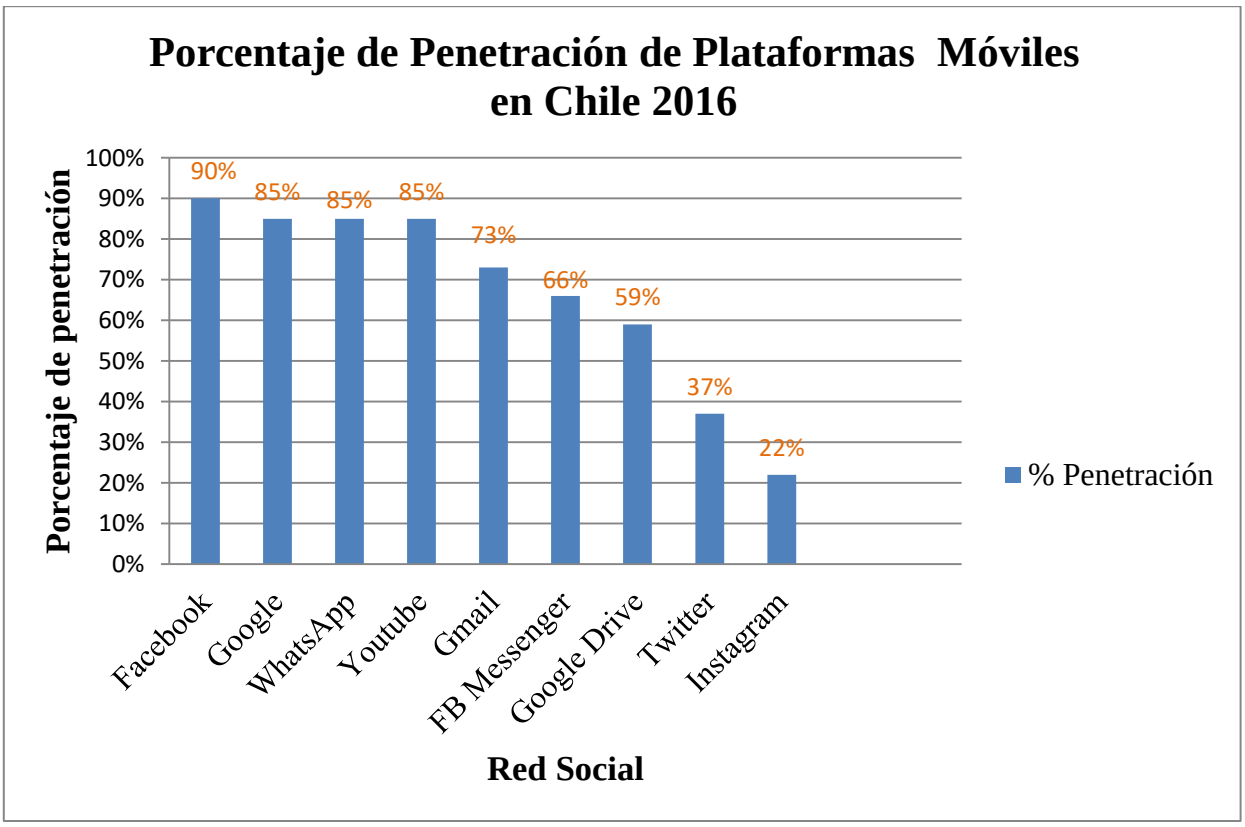

Figura 1.3: Porcentaje de penetración de plataformas móviles.

Fuente: Informe Big Data 2016, Movistar Chile [4].

El estudio considera la penetración considerando usuarios únicos conectados a las plataformas móviles a través de sus smartphones, considerando únicamente el uso de la plataforma, no el tiempo que dedica a cada una.

Al observar las plataformas móviles se observa que los sitios de redes sociales son los que se presentan con mayor participación, liderando el ranking Facebook con 90% de participación, más atrás está WhatsApp con 85% y más atrás Twitter con 37%.

Twitter está presente entre las 9 plataformas con mayor penetración, si bien no lidera el ranking cuenta con un 37% de penetración que permite considerarla como una buena fuente información para estudiar fenómenos sociales.

1.3 Los delitos en Chile.

El delito en cualquier ciudad es una problemática social que se busca eliminar lo mayor posible, de hecho distintas instituciones gubernamentales están dedicadas a esta labor, como Carabineros de Chile, Policía de Investigaciones o Paz Ciudadana.

Carabineros de Chile atiende a la población por cuadrantes, de tal manera que todo Santiago está segmentado en cuadrantes de atención.

La institución recibe denuncias realizadas por víctimas de delitos, las cuales permiten a la institución cuantificar los delitos en el sector y disponer servicios policiales en los lugares más afectados.

Carabineros de Chile genera indicadores históricos de los delitos y calcula una tasa promedio de cada tipo, de tal manera que cuando las denuncias hacen que ese indicador sobrepase cierto rango, entonces indica que se debe poner mayor atención y operar de manera especial en esa zona o con esa categoría de delito para evitar que se siga masificando.

Todas las denuncias son almacenadas en un sistema de información al cual tienen acceso cada comisaría a través del sistema AUPOL (automatización policial).

Las denuncias además de ser una fuente de datos que permite cuantificar el delito y accionar respecto a eso, permiten también analizar cómo se están realizando los delitos y cómo evolucionan.

La ley N°20.285 de Transparencia de la Función Pública y de Acceso a la Información de los Órganos del Estado faculta a cualquier persona natural a solicitar información de las denuncias.

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 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) [5].

Los DMCS son analizados y estudiados por el Departamento de Análisis Criminal de Carabineros de Chile (DAC) el cual utiliza la Plataforma de Análisis Criminal Integrado de Carabineros (PACIC) la cual les permite identificar las fluctuaciones de los indicadores de delitos y actuar ante anomalías.

El 95% de las denuncias son recibidas por Carabineros de Chile, el 5% restantes son realizadas en Policía de Investigaciones (PDI) en su brigada especializada de robos, por tanto la información que esta institución tiene es marginal comparada a la fuente de información que almacena Carabineros de Chile.

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) 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.

Dado que la ENUSC mide indicadores a través de encuestas en la población, permite cuantificar la cifra negra de las denuncias.

En la ENUSC realizada el año 2015 se cuantificaron cifras negras de algunos de los delitos de la categoría DMCS:

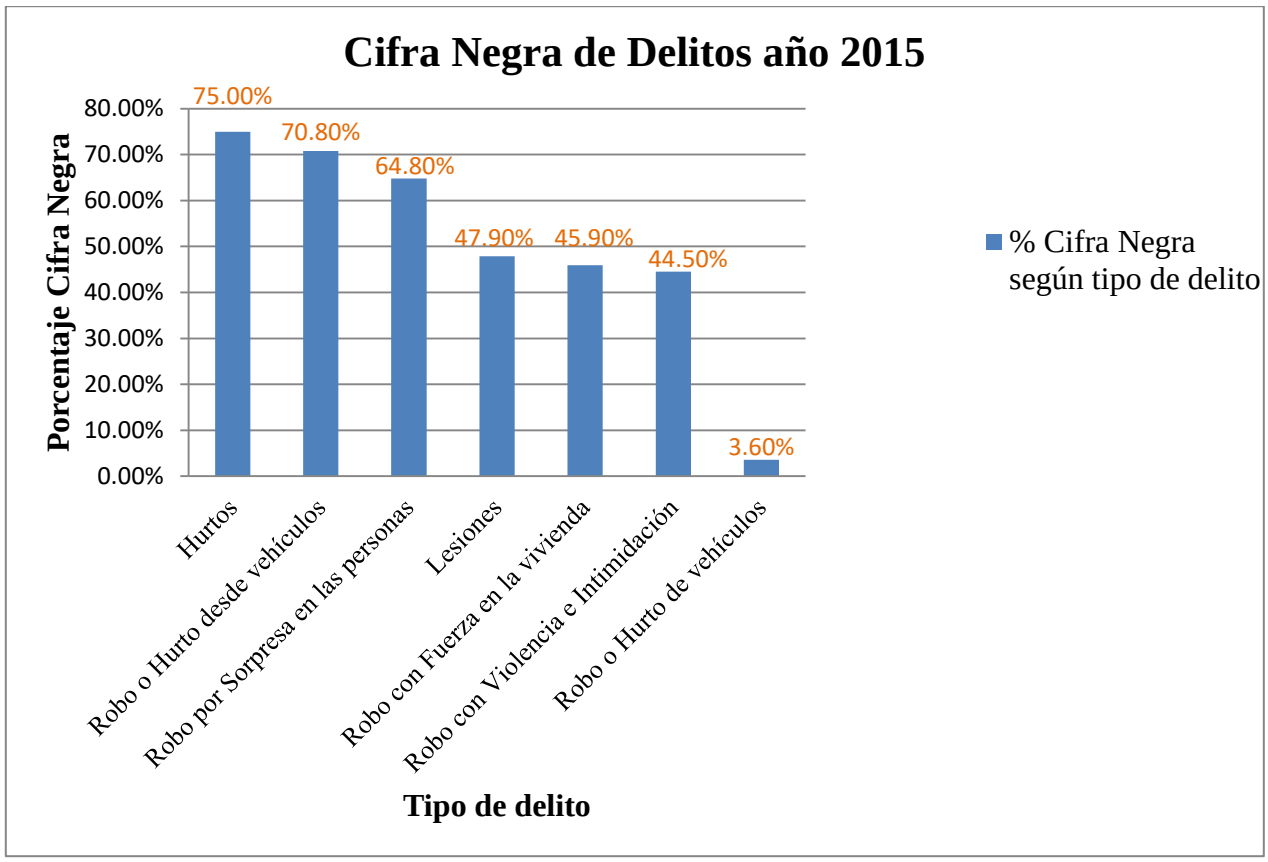

Figura 1.4: Cifra Negra de Delitos año 2015.
Fuente: Elaboración propia con datos ENUSC [6].

Como se muestra en la gráfica, el delito que menor tasa de cifra negra tiene es el "Robo o Hurto de vehículos", por lo tanto solo con ver las denuncias es posible identificar muy bien la cantidad de delitos cometidos en esta categoría y como estos varían ya que son muy representativas de la realidad del delito de robo o hurto de vehículos, en cambio la categoría "Hurtos" presenta una cifra negra del 75% lo que indica que no sería confiable la base de datos de las denuncias ya que están muy sesgadas en cuanto a cifras, hay un 75% de los delitos cometidos en esta categoría que no están siendo denunciados formalmente transformándolos en cifra negra.

1.3.1 El robo de vehículos.

El delito de robos de vehículos presenta bastante confiabilidad en cuanto al uso de las denuncias como indicador del número de delitos cometidos en esta categoría ya que la cifra negra de las denuncias de esta categoría de delito es del 3,6%.

A continuación se presenta el número de denuncias de los últimos 5 años.

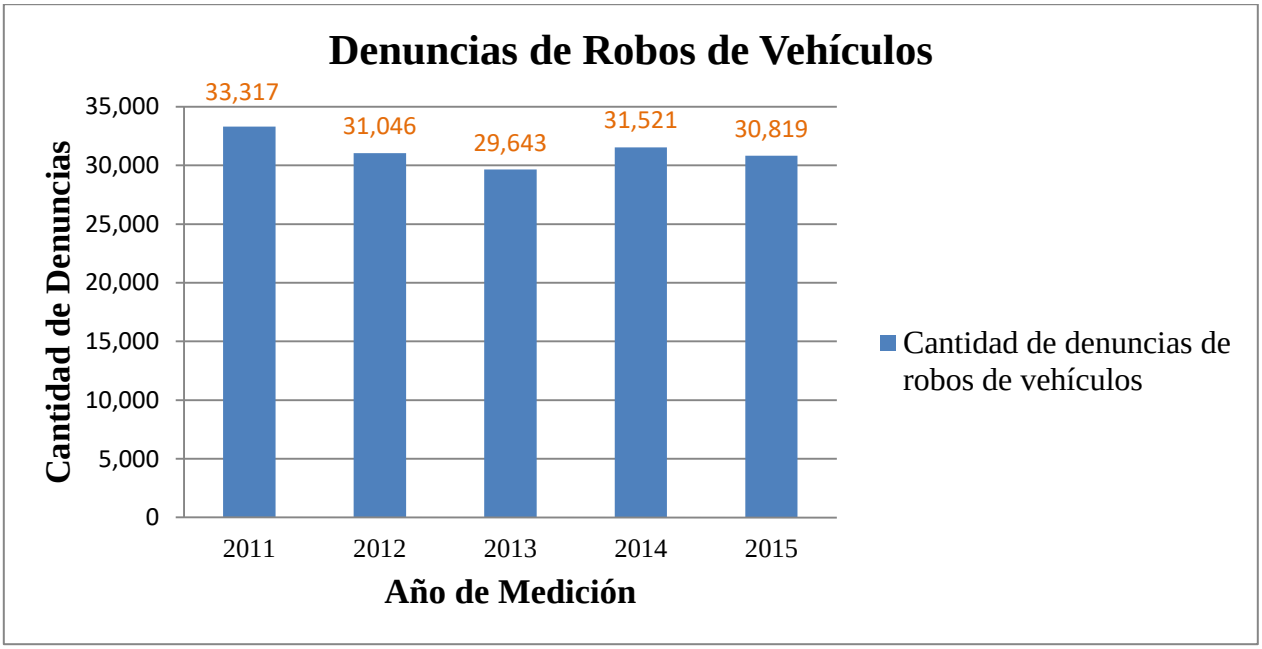

Figura 1.5: Denuncias de Robos de Vehículos.
Fuente: Elaboración propia con datos de "Estadísticas delitos de mayor connotación social" [7].

Por otra parte el número de vehículos en circulación en chile ha ido en constante crecimiento, tal como se muestra en el siguiente gráfico:

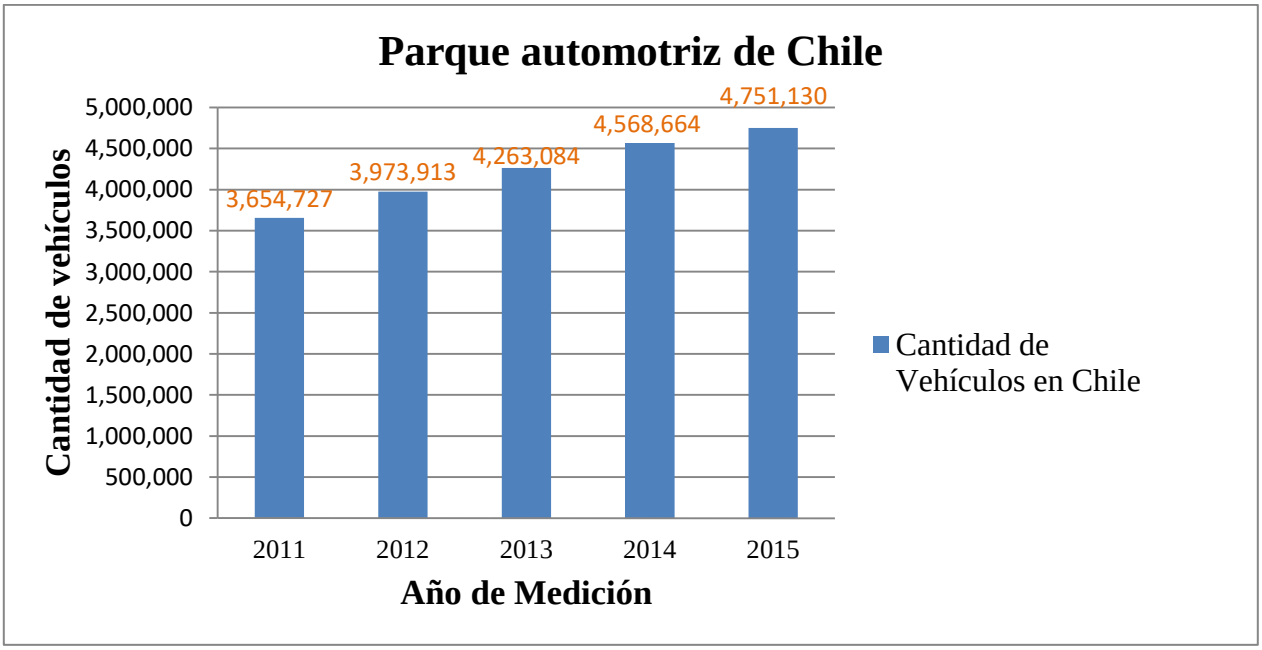

Figura 1.6: Parque automotriz de Chile.
Fuente: elaboración propia con dato de "INE ANUARIOS PARQUE DE VEHÍCULOS EN CIRCULACIÓN 2015" [8]

Lo primero que se puede entender de los gráficos es que a pesar de que el número de vehículos en circulación, por ende candidatos a ser sustraídos, ha aumentado, el número de robo de vehículos no sigue un crecimiento proporcional, sino que más bien mantiene un estabilidad en torno a los 30.000 vehículos anuales.

1.4 Objetivos.

1.4.1 Objetivo General.

El "Proyecto Observatorio del fenómeno de Robo de vehículos en Chile", pretende entender y desarrollar un modelo predictivo de los robos de vehículos que se desarrollan en Chile. Para realizar esto consolida distintas fuentes de datos, la principal son las denuncias de robos de vehículos realizadas formalmente (PROSE), y como fuentes de datos secundarias a Twitter y medios noticiosos.

Las denuncias de robos de vehículos son altamente confiables en el sentido que tienen una cifra negra menor al 4%, por otro lado Twitter es una fuente de información valiosa para los modelos predictivos al ser datos que se emiten en tiempo real, y tal como lo muestra la investigación número 3 expuesta en 2.2.1, cuando se combina con una fuente de datos principal de denuncias, el modelo puede mejorar su rendimiento.

Además tal como mostraron investigaciones anteriores realizadas en el ámbito criminal (ver 2.2.2), es importante evidenciar cuales son los sesgos inmersos en las fuentes de datos secundarias, Twitter específicamente, ya que el desconocer los sesgos de esta fuente de datos dificulta el entendimiento de patrones, empeora los resultados obtenidos o mejor dicho, el saber los sesgos permitiría reparar errores, mejorando el rendimiento del modelo predictivo. Es por esto que el objetivo general de esta tesis es:

- Identificar las relaciones existentes entre el robo de vehículos y denuncias realizadas a través de Twitter en el periodo 2012 al 2016 con el fin de establecer las diferencias y similitudes entre ambas fuentes de datos.

1.4.2 Objetivos Específicos.

Para cumplir con el objetivo principal se enuncian objetivos específicos que permiten dirigir el estudio hacia el cumplimiento del objetivo principal.

Los objetivos Específicos son:

- Identificar las metodologías aplicadas y los resultados obtenidos en investigaciones ya realizadas que hayan utilizado Twitter como fuente de información en el análisis delictual.

- Investigar y aplicar métodos de extracción de Tweets y filtro para obtener una base de datos de Twitter.

- Aplicar herramientas de minería de texto sobre la base de datos que contiene los Tweets con el fin de extraer información relevante.

- Evidenciar los sesgos encontrados en Twitter al compararlo con PROSE.

1.5 Hipótesis de investigación.

Basado en la investigación existente relacionada con el uso de twitter como una fuente de información que permite explicar fenómenos sociales es que nace la hipótesis de esta investigación definida como:

"Existe relación entre las denuncias formales realizadas por robo de vehículos y las denuncias realizadas por twitter".

1.6 Metodología.

Para realizar los objetivos detallados anteriormente se aplica primero el estudio del arte respecto al robo de vehículos identificando metodologías utilizadas que guiaran este trabajo para luego aplicar KDD, knowledge discovery in databases.

KDD es el proceso general para descubrir conocimientos útiles a través de los datos [9]. El proceso es iterativo e interactivo con muchas decisiones tomadas por el investigador.

Pasos de aplicación de KDD:

1. Entender el dominio de la investigación en donde se está aplicando KDD, cual es el problema a resolver y cuáles son sus objetivos.

2. Seleccionar del conjunto de datos originales, un subconjunto apropiado para el problema que deseamos resolver, eliminando variables irrelevantes.

3. Limpieza y pre procesamiento, decisiones sobre valores faltantes, atípicos, erróneos, ruido.

4. Reducción de los datos y proyección, encontrando los atributos útiles que permitan representar los datos de manera reducida, disminuyendo las dimensiones o transformando los datos.

5. Minería de datos, escoger que modelos utilizar en los datos considerando el objetivo declarado en el paso 1.

6. Interpretar los patrones, y posiblemente incluir visualización de los patrones obtenidos.

La iteración en el proceso KDD remite al hecho de que es común que una vez obtenido un descubrimiento de conocimiento sea necesario iterar sobre pasos anteriores para obtener nuevos descubrimientos o precisar el obtenido.

Es interactivo porque durante el proceso es necesario interactuar con diferentes expertos en el conocimiento de los datos o en el fenómeno estudiado para validar hipótesis o el conocimiento obtenido.

1.7 Resultados esperados.

Se espera desarrollar una herramienta que permita descargar Tweets.

Se espera también obtener una base de datos de Tweets que contenga denuncias de vehículos robados y obtener de dicha base de datos un vector con información de los vehículos.

Identificar sesgos en los modelos de vehículos denunciados en la red social.

Encontrar correlación en las tasas de denuncias en diferentes espacios temporales.

Identificar patrones de comportamientos no visualizados actualmente.

Capítulo 2: Marco conceptual

2.1 Twitter.

Twitter 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.

Según Twitter "500 millones de Tweets son enviados por día" [10], esto da cuenta de la cantidad de datos que son almacenados en esta plataforma de mensajes compartidos.

Twitter, como fuente de datos, presenta cualidades ¿que son de gran interés para investigadores,

Twitter presenta cualidades que hacen que se le observe como una fuente valiosa de información, algunas de las características que presenta es el hecho de que los mensajes son emitidos "en línea", es decir un ves que son emitidos por los usuarios, estos son publicados inmediatamente en la red social. Esto genera grandes ventajas frente a otras fuentes de información ya que permite ir midiendo como se va comportando un fenómeno social minuto a minuto mediante el análisis de los mensajes publicados.

Otra de las ventajas es el soporte para desarrolladores, por ejemplo ofrece tres API's, las cuales básicamente cumplen la función de facilitar el acceso a la base de datos de Twitter, de esta manera quienes estén interesados en utilizar la información publicada lo pueden hacer de manera relativamente sencilla por medio de estas herramientas.

Como se mencionó, en Twitter son publicados una enorme cantidad de mensajes diariamente, esto hace que la información obtenible para desarrolladores sea de gran interés para su análisis.

Las características de la plataforma han motivado que muchos investigadores tengan gran interés en utilizar la información publicada ahí para utilizarla como una fuente de datos que apoye sus investigaciones en diversas temáticas.

2.1.1 Tweet.

El contenido de los mensajes compartidos por los usuarios de la plataforma Twitter es conocido como Tweets el cual debe ser estructurado en no más de 140 caracteres y es transmitido en tiempo real sin costo para el usuario.

Si bien son solo 140 caracteres los disponibles para transmitir el mensaje, también se pueden adjuntar links de fotos, videos, GIFs y encuestas los cuales no se consideran dentro de los 140 caracteres.

Un ejemplo de un Tweet es el siguiente:

Figura 2.1: Ejemplo Tweet

Fuente: Twitter. [11]

El Tweet considera los siguientes campos que son posibles de extraer:

-Usuario: Nombre del usuario creador del Tweet.

-Fecha: Fecha y hora en la cual fue elaborado el Tweet.

-Contenido: El contenido expresado en el Tweet contemplado dentro de los 140 caracteres máximos.

-Acciones: Los Tweets tienen acciones ejecutables por otros usuarios que visualizan dicho Tweet, estas acciones son: Reply, ReTweet y Like. La primera contiene el número de respuestas que tuvo dicho Tweet, el segundo es la cantidad de veces que fue compartido el Tweet por otros usuarios y el tercero corresponde a la cantidad de usuarios que expresaron gustarle el contenido del Tweet.

- Geolocalización: Ubicación GPS, desde donde es emitido el Tweet, y por ende es posible de rastrear desde que 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 solo con los otros campos.

Dado que Twitter dispone de una gran cantidad de opiniones y contenido compartido por diferentes usuarios referente a diferentes tópicos es que ha motivado su uso para diferentes investigación, algunas de las cuales han mostrado que analizando el contenido de los Tweets con

respecto a un tópico en particular ha permitido desarrollar predicciones, como por ejemplo, elecciones políticas [12] o brotes de infecciones [13].

2.2 Trabajos relacionados.

2.2.1 Investigación del crimen usando twitter.

Variados trabajos destinados a la investigación del crimen usando twitter han mostrado que esta red social se puede utilizar para describir o predecir comportamientos criminales en las ciudades.

Por ejemplo una de las áreas más desarrolladas es la elaboración de Hotspots, la cual consiste en analizar mensajes publicados en Twitter relacionados con crímenes con la finalidad de establecer zonas de alta concentración de hechos criminales. Otras investigaciones han contemplado el desarrollo de modelos predictivos en esta misma área.

A continuación se presentan diferentes investigaciones desarrolladas en el área criminal:

1) Identifying Crime Hotspots using Twitter, 2015 [14]

Pregunta de investigación:

No aparece explícita en el documento, sin embargo se identifica como posible pregunta de investigación la siguiente:

¿Hay correlación entre reportes actuales de crimen y referencias de delitos publicados en Twitter?

Metodología:

Realizaron una recopilación inicial de 1 mes de Tweets, para comenzar el proyecto, y poder trabajar con datos de prueba, luego con el desarrollo del proyecto, fueron ampliando la recopilación de Tweets durante varios meses (enero a abril 2014).

Los autores analizan el lenguaje utilizado en los Tweets para términos o frases específicas que están relacionadas con el crimen.

Luego utilizando los metadatos de localización que vienen vinculados con los Tweets, 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 un entorno de software basado en la web, específicamente la API de Google Maps para trazar ubicaciones de dichos Tweets. Por otro lado trazaron en un mapa los delitos reales, información obtenida por la Policía, ambos set 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, daños criminales, 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.

Realizaron diferentes cálculos estadísticos:

Test de correlación:

Un test de correlación entre ambas fuentes de datos, enfocado solo en la cantidad de datos denunciados, es decir la frecuencia con la que se realizaban las denuncias de crímenes en ambas fuentes de datos.

El test rechazó la hipótesis nula de correlación entre ambas fuentes de datos en todos los tipos de delitos, por lo tanto no hay correlación en las tasas de frecuencia.

Calcularon la densidad de frecuencia de los delitos y los rankeaban de tal manera que al dibujarlas en el mapa, se identificaron con un color rojo intenso aquellas zonas con índices más altos en el ranking y con color amarillo aquellas que estaban al final.

Resultados:

La investigación obtuvo variados resultados, a continuación se presentan los resultados principales.

Sub-estimación de la cantidad de delitos:

En términos generales Twitter sub-estimaba considerablemente la cantidad de delitos, como en la categoría de "Comportamiento anti-social", la cual tenía una de las tasas más altas de denuncias, y 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 de la categoría "Orden público", los resultados mostraron comportamientos similares, pero con mayor magnitud de cobertura en Twitter, debido a que las protestas abarcan gran parte de área de una ciudad y por ende ampliaban las áreas de denuncias, además está el efecto del deporte en donde usan el concepto "fight" o pelea entre equipos deportivos, pero no en términos de crimen, sin embargo el algoritmo utilizado lo consideraba como problemas de orden público aumentando los puntos de alta densidad.

Error en la categorización:

En algunos casos como en la categoría de crimen "Posesión de armas", incurrían en un error al categorizar los Tweets en esta categoría ya que al hablar de juegos de armas por Twitter, el sistema lo catalogaba como una denuncia de posesión de armas y sobre estimaba la cantidad real de delitos.

Problemas definiendo la categoría:

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, de hecho no se identifica ninguna zona con alta densidad en el mapa.

Conclusiones del autor:

Una de las cosas que el autor harían mejor, sería considerar un set de testeo mucho mayor, principalmente para mejorar el mecanismo de cómo categorizar los delitos, ya que esto era fundamental para su trabajo, y con la fase de testeo se entrenó el modelo que categorizaba los Tweets.

Conclusiones propias:

La investigación mostró que Twitter en general sub estimaba las tasas de delitos, por ende es importante poder cuantificar cuánto es esta sub estimación, pensando en un modelo predictivo.

Otro tema que menciona, tal como en otras investigaciones, es la importancia de discriminar o categorizar de manera correcta los Tweets, 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 Tweets en esta categoría, y como mencionan los autores la justificación es que les fue muy difícil poder identificar cuales Tweets correspondían a esta categoría.

2) The impact of using social media data in crime rate calculations: shifting hot spots and changing spatial patterns, 2015 [15].

Pregunta de investigación:

1) Son los Hotspots de crimen estables bajo la aplicación de diferentes mediciones de tipos de poblaciones en riesgo.

2) Cuales áreas tienen la tasa de crimen más altas cuando se usan ambas poblaciones, residencial (CENSO) y mobile (Twitter) como datos de riesgo.

Metodología:

En esta investigación están interesados en el riesgo de victimización criminal desde una dimensión espacial, es decir dependiendo de su localización. Para realizar el estudio solo incluyeron mensajes en Twitter que tienen coordenadas GPS incorporadas. Dichos datos son

generados comúnmente por dispositivos móviles en donde los usuarios han optado explícitamente por publicar su localización.

La recopilación de Tweets fue realizada durante el 22 de junio del 2011 hasta el 14 de abril del 2013.

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.

En esta investigación utilizan datos provenientes de las personas, en específico, mensajes de dispositivos móviles que son publicados en Twitter. Este tipo de datos tiene el potencial de representar el ambiente de la población en una mejor disposición espacial y temporal que previas investigaciones de análisis de localización del crimen en donde utilizan la población residencial para su estimación.

Lo primero que realizan es utilizar la densidad de mensajes para marcar en un mapa las densidades de los mensajes y poder identificar rápidamente las zonas con más y menos cantidad de delitos.

Luego comparan los resultados de ambas fuentes de datos en un mapa.

Resultados:

En general los resultados no muestran correlación en los resultados al utilizar la población móvil y los resultados al usar la población residencial.

Los resultados más llamativos son los de Leeds city center, zona que tiene una 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. En consecuencia, a pesar del alto volumen de eventos criminales, no hay un crecimiento estadísticamente significativo en el riesgo de victimización criminal violenta cuando se considera la población teóricamente informada (población móvil) como población en riesgo.

Además hay un pequeño número de barrios muy cerca del centro de la ciudad que exhiben tasas de criminalidad violenta significativamente altas al considerar ambas poblaciones en riesgo, independiente del método de segmentación utilizado. No hay una razón obvia para tales tasas.

Conclusiones del autor:

Una de las conclusiones enunciadas es que deben ser precavidos con el uso de los datos de Twitter y en hacer generalizaciones sobre la población en movimiento. ¿Qué tan bien la localización espacial de los mensajes de medios sociales reflejan la localización actual de la población ambiente, en general? Se sabe que algunos grupos socioeconómicos están sobre representados en esos datos, pero ¿es esto necesariamente un problema? Además, ¿hasta qué punto (usuarios de Twitter) sesgaban la distribución espacial de la población en riesgo? Estos usuarios pueden simplemente twitear en lugares donde hay más población de todos modos, sin causar ningún sesgo espacial, o ellos podrían hacer que esta parezca como si más personas estuvieran presentes de los que realmente están presentes. 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 general, hay problemas potenciales que deben ser investigados para el uso apropiado de datos provenientes de la multitud. Sin embargo, si ellos pueden ser resueltos, ahí hay un gran potencial, particularmente para análisis espacial del crimen.

Conclusiones propias:

Tal como menciona la investigación uno de los problemas más importantes que tuvieron es el hecho de que el 1%-2% de los Tweets incorporan la ubicación GPS y esto reduce el tamaño de la fuente de datos obtenida de la recopilación de Tweets.

Un segundo tema a considerar son los sesgos que introduce la plataforma Twitter, los cuales no fueron identificados en esta investigación. Por ende descubrir estos sesgos, aporta a solucionar los problemas que presenta utilizar esta plataforma como fuente de datos en un modelo predictivo.

3) Predicting Crime Using Twitter and Kernel Density Estimation, 2014 [16]

Pregunta de investigación:

La pregunta que motiva la investigación en este caso es responder si ¿es posible usar las publicaciones de Tweets realizadas por los residentes de una de las más grandes ciudades de Estados Unidos para predecir Actividad criminal local?

Metodología:

Recopilaron Tweets durante 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 calcular la predicción de crímenes fue KDE, una técnica que ajusta un espacio bi-dimensional 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 los datos de Twitter.

Usaron una técnica de text mining para separar el Tweet en palabras o tokens. Luego analizando esas palabras identificaron el tópico del Tweets, luego se le asigna un tipo de crimen dependiendo de los tópicos identificados en el Tweet.

Para medir el rendimiento de los modelos, se grafica la capacidad de predicción del modelo, es decir se calcula AUC (AREA UNDER CURVE). Un mejor rendimiento de la predicción es indicado mediante curvas que se aproximan hacia la esquina superior izquierda del área del gráfico, por lo tanto un mayor indicador AUC significa mejor rendimiento en la predicción.

Resultados:

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 Tweets 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 porque algunos tipos de crímenes tienen mejores resultados que otros.

Conclusiones del autor:

Se concluyen al analizar los resultados que los crímenes que aumentaron más su rendimiento fueron aquellos que fueron identificados con más tópicos, es decir aquellos en que el Tweet fue mejor catalogado en el tipo de crimen.

Ellos sugieren que aumentando los tópicos posibles para cada crimen, mejorará la predicción.

Conclusiones propias:

La investigación muestra que el utilizar datos de Twitter en el modelo predictivo, en general mejora el rendimiento del modelo, en el caso particular del robo de vehículos el rendimiento es mejor al añadir los datos de Twitter al modelo.

Mencionan además, tal como las investigaciones enunciadas anteriormente, que es muy importante como se catalogan los Tweets, es decir que el código utilizado para describir el Tweet en el crimen correcto es fundamental para obtener mejores resultados en el modelo predictivo.

2.2.1.1 Conclusión de la investigación bibliográfica.

Diferentes estudios e investigaciones han sido realizados en torno al tema central del crimen en las ciudades.

Se acaban de presentar tres que exponen la manera en que se trabajan estas investigaciones, sus metodologías, resultados y conclusiones.

Todas coinciden en la importancia que tiene el código que categoriza los Tweets en los tipos de delitos, su mala identificación impacta directamente en los resultados, provocando incluso que ciertos tipos de delitos no muestren resultados debido a su mala categorización.

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, ya que al evaluar los resultados o patrones de comportamientos muchos no pueden ser entendidos debido a que no saben cuanta responsabilidad de las anomalías presentadas en los resultados pueden ser debidos a sesgos que están presentes en esta fuente de datos alternativa la cual se compone de publicaciones realizadas por diferentes usuarios pero que no considera el total de denuncias o población total expuesta a los delitos, en sus diferentes categorías. No pueden tampoco identificar cuanto 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 Tweets, se concluye que los modelos predictivos mejoran su rendimiento, su capacidad predictiva mejora significativamente al incorporar a los Tweets 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 la otra fuente de datos principal, mejorando la capacidad predictiva de los modelos.

Es así que dada estas conclusiones, la motivación de esta tesis 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.

Capítulo 3: Desarrollo.

3.1 Extracción de datos.

3.1.1 Mecanismos de extracción.

La extracción de Tweets es posible de realizar de manera gratuita con mecanismos de conexión con la base de datos de twitter, en donde la dificultad se dirige hacia hacer los filtros correspondientes para poder captar los Tweets de interés, en este caso los relacionados con robos de vehículos en Chile.

Para la extracción de Tweets 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 [17].

REST API (históricos): Permite extraer datos de twitters históricos [18].

Streaming API: Permite obtener Tweets en tiempo real a través de fijar palabras filtros, todos los Tweets escritos que contengan las palabras filtros serán recibidos por la API [19].

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 solo el texto, todo el HTML de la página o incluso las imágenes dispuestas en la misma. La única condicionante para usar este mecanismo es entregar una página web a la cual acceder y por lo tanto no es exclusiva para la página web de Twitter, puede ser utilizada en cualquier página web que sea permitido acceder.

3.1.2 Elección del mecanismo de extracción.

Para efectos del observatorio es necesario obtener datos en tiempo real y por lo tanto utilizar la Streaming API que permite ir recibiendo todos los Tweets que se vayan originando cuyo contenido tenga las palabras filtros fijadas por el observatorio.

Para el caso de esta investigación no es aplicable la Streaming API ya que el propósito de esta investigación es relacionar la información contenida en twitter con la base de datos de PROSE la cual considera un periodo de tiempo desde el año 2012 hasta el año 2016. Si se utilizara Streaming API para crear la base de datos de twitter, se estaría perdiendo mucha información e incluso los resultados podrían no ser claros ya que no se podría analizar el comportamiento histórico sino que de tan solo unos meses.

Descartada Streaming API, lo ideal es utilizar REST API o el mecanismo de Web Scraping.

REST APIs: permite acceder de manera programada para leer y escribir datos en Twitter. Crear nuevos Tweet, leer perfil de usuarios, datos de seguidores y más.

El problema de REST APIs es que no está creada para hacer una descarga masiva de Tweets y por ende con esta API no es posible obtener una base de datos de Tweets históricos.

Web Scraping: permite acceder a la página de twitter.com, escribir en su campo de búsqueda los términos que se deseen filtrar junto con el rango de fecha sobre la cual se quiere realizar la consulta.

Con esta herramienta se puede obtener una recopilación de Tweets que cumplan las condiciones de palabras filtros y en la periodicidad especificada los cuales en su conjunto generan una nueva base de datos para esta investigación.

Existe un pequeño problema al utilizar esta metodología, y es que al utilizar el campo de búsqueda, Twitter omite algunos Tweets ya que este considera que el contenido es catalogado como spam (mensajes no deseados) o mensajes irrelevantes ya que el usuario no tiene un comportamiento activo. Si bien esta característica produce que menos mensajes sean desplegados como resultados, también ayuda a evitar considerar Tweets repetidos que son catalogados como spam.

Dadas las opciones mencionadas anteriormente para la extracción de Tweets, se decide utilizar el mecanismo de Web Scraping para obtener una base masiva de los resultados de búsqueda pública de Twitter, principalmente porque es el único que permite acceder a Tweets históricos en el periodo 2012 – 2016.

3.1.3 Extracción de Tweets con la herramienta seleccionada.

Para realizar el Web Scraping sobre la página web a la cual se desea acceder, en este caso Twitter, es recomendable leer el archivo robots.txt [20], el cual señala las directrices que deben seguir los buscadores como google o yahooo al momento de buscar información en Twitter.com. Dentro de las direcciones (site links) disponibles a acceder que se enuncian en ese archivo está el acceso al campo de búsqueda público de Twitter [21], que no requiere identificarte con un acceso de cuenta de usuario de Twitter para poder ser utilizado.

El código de Web Scraping simula un navegador, el cual es conducido por códigos a través del software R, y de esta manera el navegador tiene un funcionamiento automatizado.

Una vez que el navegador ha ingresado a la dirección solicitada, se ingresa al campo de búsqueda de Twitter en donde hay que especificar qué términos desean ser filtrados.

Se escogieron los siguientes términos de búsqueda:

robo patente OR robado patente OR robaron patente

La razón de haber escogido esos términos es porque se visualizaron diferentes Tweets relacionados con robo de vehículos, y en todos aquellos en que se denunciaba un robo de vehículo, se mencionaba la patente.

El objetivo de la extracción de Tweets 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 PROSE.

Al utilizar los términos escogidos se podrá luego extraer de los Tweets 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 PROSE.

La patente no solo 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 Tweets, los cuales presentan dos características que limitan la extracción de información correspondiente a las características de los vehículos.

La primera es la acotación del Tweet a 140 caracteres, lo que provoca que las denuncias hagan referencias principalmente a características 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 Aplicación KDD

3.2.1 Dominio de la investigación y problema a resolver

La investigación se desarrolla en el tópico del robo o hurto de vehículos en Chile.

El robo de vehículos en Chile presenta tasas bastantes altas, 80 vehículos aproximadamente son sustraídos diariamente en el país, por otra parte el 30% del parque automotriz cuenta con seguro particular, los cuales cubren principalmente daños y robo del vehículo, de las dos coberturas, el robo de vehículos es el que involucran más costos para la aseguradora y es por eso que es de gran interés para estas compañías el poder evitar estos delitos. Además existe el interés natural por parte de la sociedad de erradicar o disminuir lo mayor posible los actos delictuales.

Con el interés en el estudio del robo de vehículos se desarrolla el “Proyecto Observatorio del fenómeno de Robo de vehículos en Chile” el cual desarrollará un modelo predictivo utilizando diferentes fuentes de datos.

Según investigaciones desarrolladas en el ámbito criminal Twitter mejora el rendimiento de los modelos predictivos que utilizan como fuente de datos principal las denuncias de los delitos, sin embargo Twitter presenta sesgos propios de una plataforma de usuarios que sobre representa a cierto grupo, sin ser evidentes estos sesgos, es importante descubrirlos, ya que favorece la interpretación de patrones de comportamientos que muestren los datos al estudiar el evento delictual. Además el descubrir los sesgos de la fuente de datos permitirá corregir errores en el modelo predictivo, ajustándolo y mejorando su rendimiento.

Por lo tanto como trabajo previo al modelo predictivo del proyecto en el que se enmarca esta investigación, nace el problema de desconocer los sesgos o las relaciones existentes entre ambas fuentes de datos.

3.2.2 Selección de datos

Bases de datos:

Base de datos de PROSE:

Contexto:

La base de datos de PROSE consolida la información referente a las denuncias de robos de vehículos realizadas por las personas que sufrieron la sustracción de su vehículo y que además cuentan con seguro particular, por lo que los vehículos propensos a estar en esta base de datos no considera el universo total del parque automotriz sino que el 30% aproximadamente que es la proporción de vehículos protegidos con algún seguro particular.

Periodo considerado:

Para el estudio de esta investigación y para nutrir de información al observatorio, PROSE ha proporcionado una base de datos que contiene las denuncias realizadas por clientes de las aseguradoras por haber sufrido el robo o hurto de su vehículo entre el periodo año 2012 hasta el año 2016.

Atributos de la base:

La base original considera 74 atributos distintos, varios de los cuales remiten a la misma información solo que usando nomenclatura distintas por ejemplo comuna, y comuna id, en el primera está escrita la comuna, en el segundo se describe con un número identificador.

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 de la hora del robo del vehículo.

-Fecha Prose: Fecha de registro del robo en PROSE.

Número de datos:

La base de datos considera 45.018 registros de robos de vehículos.

Base de datos de Twitter:

Contexto:

El observatorio incluye fuentes de datos online cuyo fin principal es nutrirlo de información complementaria a las denuncias y anticiparse a las variantes existentes en el robo de vehículos.

Una de las fuentes principales es Twitter, la cual al ser una herramienta de publicación de contenido en tiempo real permite obtener información de manera inmediata.

Para realizar la presente investigación se consideró una extracción de Tweets relacionados con el robo de vehículos, los cuales fueron obtenidos mediante el mecanismo especificado en 3.2.3.

Periodo considerado:

Para realizar la extracción de Tweets se consideró el mismo periodo contemplado en la base de datos de PROSE, es decir entre el año 2012 hasta el año 2016, ya que para analizar las relaciones entre ambas bases de datos es fundamental que los periodos de ambas bases de datos sean iguales o se perdería comparación histórica.

Atributos de la base:

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. Para poder distribuir los datos en variables es necesario aplicar minería de texto.

Ejemplo de un Tweet 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 957735036 o a carabineros al 133 RT!

Responder Retwittear 163 Me gusta 9”.

La minería de datos se aplica sobre el Tweet, 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 Tweet: “Favor RT! robo de Mazda Artis, 1999, color rojo burdeo, patente SU 3765.

Si lo ven avisen al 957735036 o a carabineros al 133 RT!”.

Reply: “”.

ReTweet: “163”.

Like: “9”.

Con el uso de las técnicas de minería de texto sobre todos los Tweets extraídos fue posible construir las variables mencionadas.

Un dato que es muy importante de obtener del texto simple del Tweet es la patente del vehículo robado que se está denunciando, la cual se encuentra en la variable “Texto Tweet”.

Mecanismo de extracción de las patentes:

Para realizar la extracción de las patentes contenidas en el texto de la variable “Texto Tweet” se aplicó minería de texto el cual se detalla a continuación.

Lo primero fue identificar visualmente que cuando se incorpora el código de la patente de un vehículo robado en la denuncia del Tweet, esta viene inmediatamente después de la palabra “patente” además la codificación de las patentes en Chile tienen 6 caracteres, pero en el Tweet puede incorporar otros caracteres como guiones o tildes, provocando que la extensión de la patente supere los 6 dígitos. Teniendo esto en consideración se procede así:

1° Se transforma todo el texto contenido en la variable “Texto Tweet” a minúscula.

2° Se ubica la palabra “patente” en el texto y se almacena el número de la posición en donde está ubicada.

3° Se almacenan los 9 caracteres posteriores de la palabra “patente”.

4° Se busca en los 9 caracteres almacenados si hay un número, ya que de no haber números entonces esta no sería una patente válida.

5° Se limpian la patente, eliminando los signos de puntuación, espacios y guiones que podrían estar contenidos.

6° Se limita la patente a solo 6 caracteres, obteniendo finalmente la patente denunciada por robo.

Realizada la minería de datos en los Tweets, se genera la base de datos de Twitter que contiene los siguientes campos:

-Usuario: Nombre del usuario que crea el Tweet.

-Fecha: Fecha de creación del Tweet.

-Hora: Hora de creación del Tweet.

-Texto Tweet: Contenido del Tweet, en el caso de esta investigación son las denuncias de robos de vehículos realizadas por la red social.

-Reply: Número de respuestas que ha tenido el Tweet.

-ReTweet: Número de veces que se ha compartido el Tweet por otros usuarios.

-Like: Número de usuarios que han indicado que les gusta el contenido del Tweet.

-Patente: Código de la patente del vehículo denunciado en el Tweet.

Numero de datos:

Una vez realizada la extracción de los Tweet desde Twitter y aplicada las técnicas de minería de textos mencionadas anteriormente se obtuvo como resultado una base de datos que contiene 18.112 registros.

3.2.3 Limpieza y pre procesamiento.

Limpieza:

Para el caso de la limpieza se procede a tratar datos inválidos o fuera de rango:

PROSE:

Registros No válidos:

De los 45.018 registros de la base de datos, 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.

Registro Recuperables:

De los 44.585 datos, 1.744 registro 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 de la Función Pública y de Acceso a la Información de los Órganos del Estado 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 de Vehículo.

-Marca.

-Modelo del Vehículo.

Con la información de Plantas de Revisión Técnica se obtuvieron los datos de:

-Color del vehículo.

Twitter:

Registros No válidos:

Anteriormente se había aplicado minería de texto sobre el Tweet para extraer las patentes denunciadas por robo y luego estos datos fueron asignados a la variable "Patente", y se había mencionado que podía haber patentes no válidas.

Ahora al analizar la variable "Patente" se evidenciaron aquellas patentes no válidas, ya sea porque no correspondía a una codificación de patente (números y letras), o porque la patente no correspondía a un vehículo chileno.

6.113 datos en la variable patente correspondían a datos no válidos.

Un ejemplo de un Tweet que no contenía una patente de vehículo:

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

En negrita se aprecia que el Tweet 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 Tweets que fueron identificados con patentes no válidas fueron eliminados ya que no correspondían a denuncias de robos de vehículos chilenos.

Eliminados los 6.113 Tweets que no corresponden a denuncias de robos de vehículos chilenos, la base de datos de Twitter permanece con 11.999 válidos.

Pre procesamiento:

Imputación de datos:

Paso importante en el pre procesamiento es la imputación de datos. La imputación de datos corresponde a la etapa final del proceso de depuración de datos, en el cual los valores faltantes ("missing") 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 [22].

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 [23].

Tipos de imputación de datos:

Reemplazo por el promedio:

Este tipo de imputación usa el promedio de los valores reportados de la misma variable para llenar los espacios vacíos.

Uno de los principales problemas con este método de imputación es que se sobre muestrea el valor promedio en la variable imputada, haciendo que no se mantenga la distribución o las relaciones multivariadas con los datos de las otras variables. Aún más puede provocar que la varianza pueda ser sustancialmente sub estimada, sin embargo es de los tipos de imputación más simples y rápido en cuanto a procesamiento.

Hot Deck:

Este método va imputando de una en una cada variable con dato vacío o faltante. Lo primero que hace el método es escoger un registro con variable faltante, luego busca un registro similar en la base de datos por medio de comparar las variables que acompañan a la imputada. Al encontrar registros similares al que está siendo imputado, se calcula el promedio de este subconjunto registros similares al que está siendo imputado y se introduce este promedio en la variable que tenía el dato faltante.

Si bien calcula el promedio, este está basado en la relación que tenga la variable imputada con las demás variables, y relacionada con registros similares y no con todos los datos, por ende es mejor que la imputación por el promedio.

Reemplazo por Regresión:

Este método de imputación usa variables auxiliares que presenten algún grado de correlación con la variable imputada y que tienen data disponible para todos o casi todos los registros.

Las variables correlacionadas son usadas para predecir el valor de la variable a través de una regresión. En este caso la capacidad de ser un buen valor imputado dependerá de que tan bueno sea el modelo escogido junto con las variables predictivas.

Al igual que en todo modelo regresivo lineal, el objetivo de la regresión es minimizar la distancia desde la recta hacia los valores de la variable a predecir.

Existe un problema cuando se trata de variables del tipo “String” o no numéricas. Ya que el modelo regresivo no puede predecir este tipo de dato. Sin embargo para este caso existe otra manera de predecir la variable calculando distancia.

Levenshtein distance:

Para definir una distancia entre variables no numéricas se puede utilizar la distancia "Levenshtein distance", la cual mide la distancia de edición que hay entre 2 palabras, por definición es el número mínimo de operaciones necesarias para transformar una cadena de caracteres en otra.

Así por ejemplo:

RAV -> RAV4 Corresponde a la inserción de un carácter.

Homologación de variables:

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.

Variables homologadas:

1) 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 solo 1 palabra, ya que ensuciaría el posterior análisis de esta variable.

Gris Metalizado -> Gris.

Codificación:

Una vez realizada la reducción a 1 palabra, la variable fue necesario modificar la etiqueta de ciertas categorías.

NARANJA -> NARANJO.

CAFÉ -> CAFE.

0 -> OTRO. (El cero hace referencia a que no se tenía información del color del vehículo).

2) 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.

KIA -> KIA MOTORS.

MITSUBISHI FUSO -> MITSUBISHI.

HI LUX -> HILUX.

RAV 4 -> RAV4.

3) 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) [24].

Categoría AUTOMOVIL:

“AUTOMÓVIL” o “AMBULANCIA” o “COCHE MORTUORIO” o “TAXI BASICO” o “LIMUSINA” o “JEEP” -> “AUTOMOVIL”.

Categoría CAMIONES:

“TRACTO CAMION” o “TRACTOCAMION” o “CAMION” o “CAMIÓN” o “CHASIS CABINADO” o “CHASISCABINADO” o “CHASIS” -> “CAMIONES”.

Categoría OTROSTIPOS:

MAQUINA INDUSTRIAL o REMOLQUE o REMOLQUES o SEMIREMOLQUE o TRACTOR o TOLVAS HIDRAULICAS (S/TOMA FUERZA) o MOTOR HOME o MAQUINA AGRICOLA o ESTANQUES METALICOS DE ACERO INOXIDABLE (SIN CHASIS) o CARROS DE ARRASTRE o BICICLETA o BICIMOTO -> OTROSTIPOS.

Categoría BUSES Y TAXIBUSES:

BUS o MICRO -> BUSES Y TAXIBUSES.

Categoría MOTO:

MOTOCICLETAS -> MOTO.

Creación de nuevas variables:

En esta parte del proceso se crearán nuevas variables que permitirán realizar un mejor análisis de las relaciones existentes entre ambas bases de datos o mejorar las conclusiones posteriores:

1) Creación variable modelo 2 string:

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 solo 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.

2) 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 crear esta variable se realizaron 2 pasos que se definen a continuación:

1- Obtener información histórica de valorización fiscal del SII:

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 [24], los cuales contienen las siguientes variables:

Tipo Vehículo – Marca – Modelo – Año – Valorización según año.

La última variable "Valorización según año" es la valorización del vehículo para el año correspondiente a la medición fiscal.

Por ejemplo la valorización de un vehículo para el año 2013:

AUTOMOVIL – FIAT – PUNTO 55 S – 2000 - $1.420.000

2- Realizar emparejamiento con las bases:

Para realizar el emparejamiento entre la tasación del SII y las bases de datos de PROSE y Twitter se realizó en dos pasos:

El primer paso fue realizar el emparejamiento sin hacer modificaciones en los datos.

Se seleccionaba un registro de las bases de datos de PROSE o Twitter con las variables:

"Tipo Vehículo - Marca - Modelo - Año - Año denuncia" y se emparejó con las variables similares en el archivo de valorización fiscal:

"Tipo Vehículo – Marca – Modelo – Año – Valorización según año".

En donde la variable "Año denuncia" hace referencia al año en que se denunció el robo del vehículo y por ende se empareja con el año de valorización fiscal, pero desfasando 1 año. La razón del desfase de un año es porque hay casos de modelos de vehículos que por ejemplo son año 2015 pero se comienza a vender en el año 2014 y por ende no aparece en el archivo de valorización del SII.

Una vez realizado el emparejamiento sin modificar los datos se obtuvieron aproximadamente el 20% de las valorizaciones de los vehículos. La razón principal es que como se mencionó anteriormente en Chile los modelos de los vehículos no están homologados y no existe una nomenclatura para ser registrados y por ende no coinciden con el archivo de valorización del SII.

Dado lo anterior es que fue necesario realizar un segundo paso.

El segundo paso para lograr obtener la valorización de los vehículos denunciados consiste en aplicar uno de los mecanismos de imputación de datos mencionados anteriormente, "Levenshtein distance".

Se realizó el nuevo proceso de emparejamiento para el 80% de datos faltantes aplicando primero una segmentación y luego usando Levenshtein distance.

La segmentación consistió en considerar un registro en la base de PROSE o Twitter con las siguientes variables:

Tipo Vehículo - Marca - Año - Año denuncia.

No se consideró el modelo, y luego se emparejaron estos datos con la información del SII, obteniendo un subconjunto de datos que son candidatos a ser emparejados con el registro al cual se le quiere agregar la valorización.

Al subconjunto de datos candidatos a ser emparejados se le calculó Levenshtein distance entre cada uno de los modelos del subconjunto y el modelo del registro seleccionado al comienzo.

Con un listado del valor de la distancia entre cada modelo se seleccionó la valorización del modelo que presentaba menor valor de distancia, reconociendo que es el más parecido en cuanto a escritura con el modelo del registro.

Aplicando Levenshtein distance se completó el 80% de tasaciones faltantes y por lo tanto el 100% de los registros de la base PROSE y Twitter obtuvieron el dato de valorización.

3) 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 PROSE como la base de datos de referencia, se calcularon los cuartiles de tasación, obteniendo:

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

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

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

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

Con las reglas para obtener los cuartiles en PROSE 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 PROSE 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 PROSE, considerando a esta base de datos como la base de datos de referencia.

3.2.4 Reducción de los datos.

Una vez ya realizado el pre procesamiento y la limpieza de los datos, se puede analizar o tomar la decisión de si es necesario reducir los datos.

La decisión depende de 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.

Corte en la base de datos:

La primera reducción de datos realizada en las bases de datos fue eliminar los vehículos 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:

Automovil - Camioneta - Furgon – Minibus - Moto - StationWagon – Todo Terreno.

La razón de haber tomado esta decisión es porque en la base de datos de PROSE, los vehículos que no correspondían a este tipo de vehículos ensuciaban mucho los datos, por ejemplo en el caso de los camiones, cerca del 50% de los modelos en esta categoría eran escrituras que no

identifican a un modelo, y al hacer los pasos previos de transformación fueron modificados por el modelo más parecido usando Levenshtein distance, pero estos modelos no representan necesariamente el modelo correcto. Además al revisar la información obtenida por medio de Ley de transparencia, también sucedía que algunos modelos de camiones o vehículos pesados eran simplemente un par de letras los cuales no son fehacientes de un modelo correcto. En el caso de los vehículos livianos esto no sucedía, las consecuencias en el caso de los vehículos no livianos son por ejemplo que las diferencias en valores entre un modelo y otro similar en escritura puede variar en un 200% por lo que se podría incurrir en grandes errores.

Además PROSE consideraba otros tipos, como "MAQUINA INDUSTRIAL" o "TOLVAS HIDRAULICAS (S/TOMA FUERZA)" que también ensuciaban los datos.

Finalmente la decisión fue tomada considerando que los vehículos no livianos representan un 9% de los datos tanto en la fuente de datos PROSE como en la fuente de datos Twitter. Por lo tanto no se pierde el foco principal de esta investigación, de hecho se focaliza mejor la investigación, permitiendo interpretan de mejor forma los resultados posteriores.

Esta reducción de datos significo:

PROSE: reducir los registros de 44.585 registros a 41.692.

Twitter: reducir los registros de 11.999 registros a 11.025.

Eliminación de filas duplicadas:

La base de datos de registros de denuncias de PROSE 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 Tweets que denuncian el mismo vehículo, esto se pudo comprobar a través de la variable patente, la cual fue extraída del Tweet. Si bien el Tweet puede no ser idéntico, están denunciando al mismo vehículo por lo que en este caso se mantuvo aquel Tweet que fue emitido antes, es decir aquel que presenta mayor antigüedad. Según este procedimiento se identificaron 3.758 Tweets que denuncian un vehículo ya denunciado anteriormente en otro Tweet.

La reducción de filas duplicadas significó:

PROSE: reducir los registros de 41.692 registros a 39.745.

Twitter: reducir los registros de 11.025 registros a 7.267.

Antes de proseguir con el proceso de minería de datos se detalla cómo han quedado establecidas ambas bases de datos con los pasos del proceso KDD ya realizados hasta acá.

Base de datos de PROSE:

39.745 registros.

Patentes – Tipo Vehículo – Marca – Modelo – Año – Color- Modelo 2 string - Tasación – Grupo Tasación – Fecha Siniestro – Fecha Prose - Fecha Denuncia – Fecha Hallazgo – Si Tweet.

A continuación se muestra un extracto de la base de datos (las patentes fueron modificadas).

Patentes	Tipo Vehículo	Marca	Modelo	Año	Color	Modelo 2 string	Tasación	Grupo. Tasación	Fecha Sin	Fecha Prose	Fecha Den	Fecha Hallaz	Si tweet
IJ1234	AUTOMOVIL	TOYOTA	YARIS GLI	2010	BLANCO	YARIS GLI	$ 5.090.000	2	27-01-2014 5:00:00	05-02-2014	27-01-2014	29-01-2014	1
JK1234	STATION WAGON	KIA MOTOR	GRAND CARNIVA	2012	BLANCO	GRAND CARNIV	$ 8.420.000	4	24-01-2014 6:30:00	03-02-2014	24-01-2014	10-02-2014	0
KL1234	AUTOMOVIL	CHEVROLET	AVEO LT HB 5P 1.4	2007	GRIS	AVEO LT	$ 2.980.000	1	10-02-2014 12:30:00	19-02-2014	10-02-2014	NA	0
LM1234	CAMIONETA	NISSAN	TERRANO DCAB DXS	2011	ROJO	TERRANO DCAB	$ 5.140.000	2	07-02-2014 0:15:00	17-02-2014	07-02-2014	NA	0
MN1234	FURGON	FIAT	FIORINO FIRE 1.2	2011	BLANCO	FIORINO FIRE	$ 4.380.000	2	22-02-2014 9:30:00	28-02-2014	23-02-2014	NA	0
NO1234	AUTOMOVIL	HYUNDAI	I30 GLS 1.6	2012	AZUL	I30 GLS	$ 9.370.000	4	24-02-2014 18:15:00	28-02-2014	25-02-2014	NA	1
OP1234	AUTOMOVIL	HYUNDAI	ACCENT GL 1.4	2011	AZUL	ACCENT GL	$ 5.110.000	2	18-02-2014 4:00:00	28-02-2014	18-02-2014	26-02-2014	0
PQ1234	AUTOMOVIL	HYUNDAI	ELANTRA GLS 1.8	2012	AZUL	ELANTRA GLS	$ 6.390.000	3	25-02-2014 4:00:00	28-02-2014	25-02-2014	21-03-2014	0

Tabla 3.1: Base de datos PROSE.

Base de datos de Twitter:

7.267 registros.

Patentes -Tipo Vehículo – Marca – Modelo – Año – Color – Modelo 2 string- Tasación – Grupo Tasación - Reply - ReTweet - Like - User - Fecha.

A continuación se muestra un extracto de la base de datos (las patentes fueron modificadas).

Patentes	Tipo Vehículo	Marca	Modelo	Año	Color	Modelo 2 string	Tasación	Grupo. Tasación	Reply	Retweet	Like	User	Fecha
AB1234	AUTOMOVIL	HONDA	CIVIC LSI 1.5	1995	VERDE	CIVIC LSI	$ 1.250.000	1	1	41	6	@Quinta_Coqu	29-12-2016 9:39:36
BC1234	CAMIONETA	GREAT WALL	DEER 2.2	2008	BLANCO	DEER 2.2	$ 2.120.000	1	0	0	0	@SilvPadilla	28-12-2016 19:13:12
CD1234	AUTOMOVIL	NISSAN	V16 SENTRA 1.6	2008	PLATEADO	V16 SENTRA	$ 2.650.000	1	0	1	1	@laamistaddet	27-12-2016 9:56:24
DE1234	AUTOMOVIL	NISSAN	V16 EX SALOON	1998	NEGRO	V16 EX	$ 1.230.000	1	0	10	3	@cbusca_cl	27-12-2016 8:42:26
EF1234	AUTOMOVIL	FIAT	DUNA SEDAN	1992	OTRO	DUNA SEDAN	$ 690.000	1	1	7	1	@CEBB_241	26-12-2016 18:28:04
FG1234	STATION WAGON	HYUNDAI	SANTA FE GLS 2.4	2011	PLATEADO	SANTA FE	$ 8.740.000	4	0	0	0	@Edo2509	22-12-2016 5:17:33
GH1234	AUTOMOVIL	HYUNDAI	ACCENT PRIME GL	2002	NEGRO	ACCENT PRIME	$ 1.780.000	1	1	1	0	@EmerArauca	24-12-2016 4:17:10
HI1234	AUTOMOVIL	NISSAN	SENTRA 1.6 EX SALOON	1993	BLANCO	SENTRA 1.6	$ 760.000	1	0	12	0	@infobiobio	23-12-2016 11:03:26

Tabla 3.2: Base de datos Twitter.

3.2.5 Minería de datos.

A continuación en el proceso de minería de datos 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.

Inspección visual de las Tasas de Frecuencia:

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.

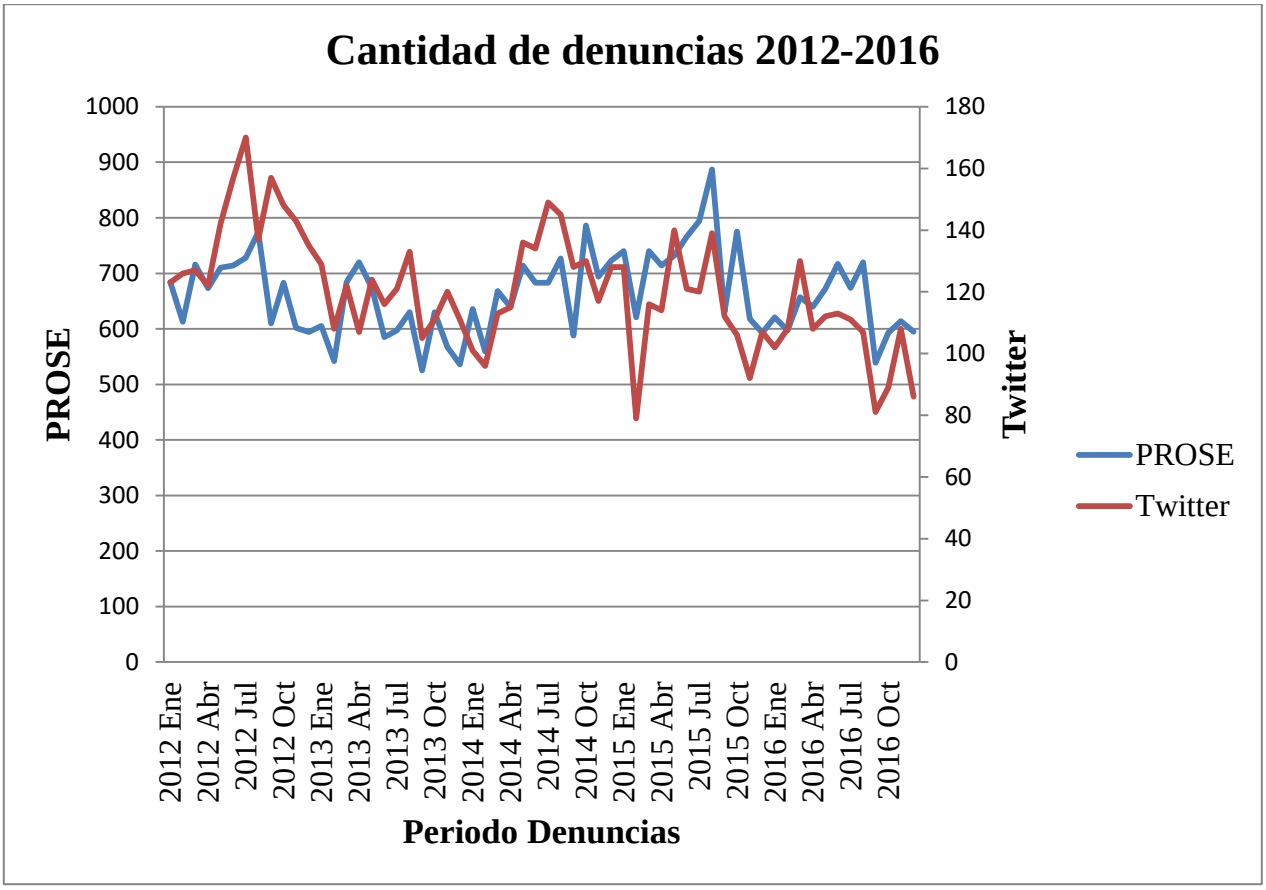

Figura 3.1: Gráfico Cantidad de denuncias 2012-2016.

En el eje Y están las cantidades de denuncias según la fuente de datos. En el eje X está el periodo considerado, es decir el mes al cual corresponde la medición.

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.

Como en el periodo 2015-2016 se aprecia una mayor correlación en cuanto al comportamiento, se realizará una nueva inspección visual de este periodo en particular.

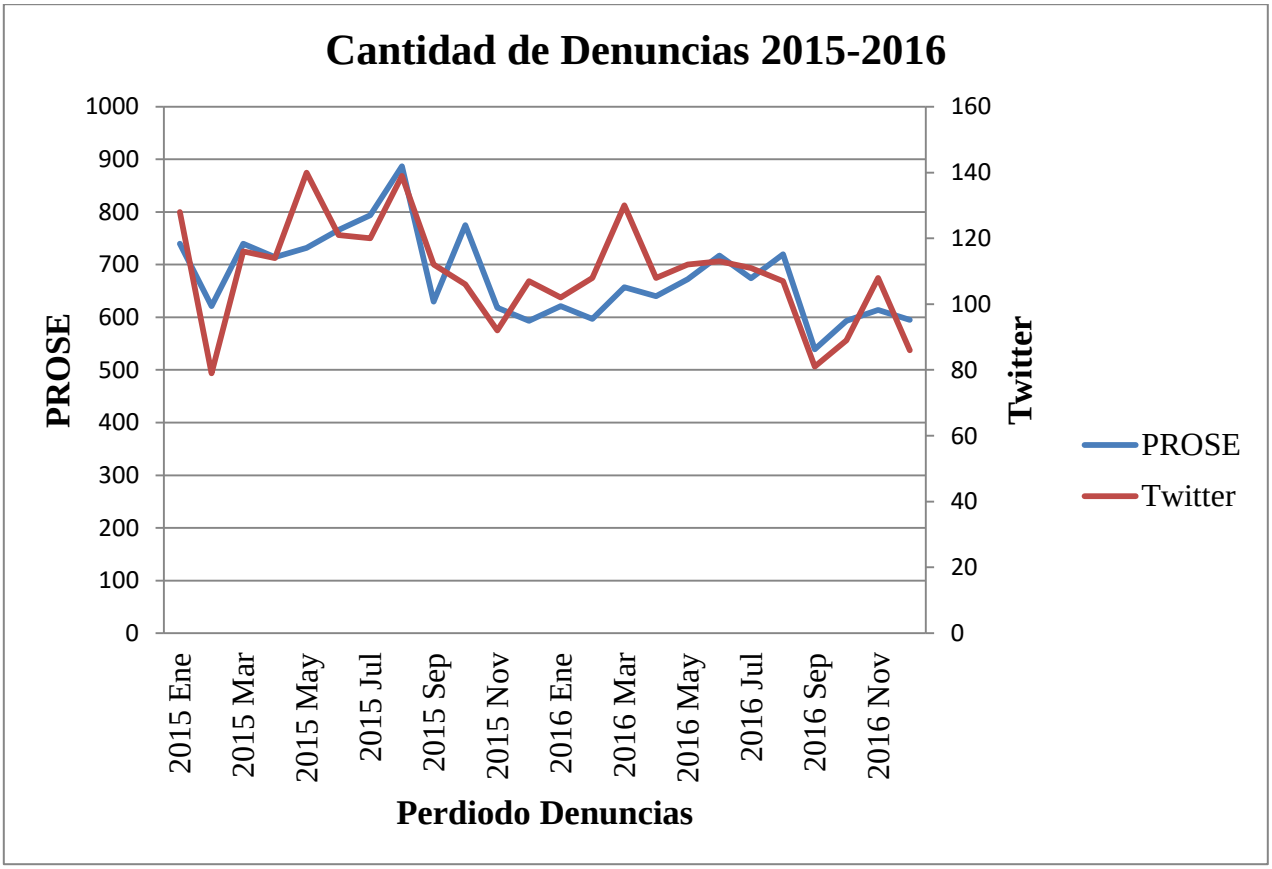

Figura 3.2: Gráfico Cantidad de denuncias 2015-2016.

En el eje Y están las cantidades de denuncias según la fuente de datos. En el eje X está el periodo considerado, es decir el mes al cual corresponde la medición.

Analizando este periodo en particular se puede evidenciar con mayor claridad que en los picos de frecuencia ambas bases de datos presentan alta correlación, se manifiestan de maneras similares.

Ya inspeccionado visualmente las tasas de frecuencias es que se calculará la correlación entre ambas bases de datos.

Coeficiente de Correlación en tasa de frecuencias:

Un buen método para entender de manera numérica el grado de relación que tienen 2 fuentes de datos distintas es aplicar un test de correlación. En este caso el coeficiente de correlación entre ambas bases de datos se realiza sobre las frecuencias de denuncias de robo de vehículos.

El coeficiente de correlación para el periodo 2012-2016 muestra: 0,42 de correlación.

Lo que muestra que correlación hay pero no es alta para decir que hay una relación directa entre ambas fuentes de datos.

Sin embargo dado que visualmente se mostraba mayor relación en el periodo 2015-2016, es que se calcula el coeficiente de correlación para este periodo.

El coeficiente de correlación para el periodo 2015-2016 muestra: 0,73 de correlación.

Este si es un coeficiente alto de correlación y muestra que hay una relación directa en la frecuencia de las denuncias de ambas bases de datos en este último periodo.

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 tipo de vehículo, grupo de tasación y modelo de vehículo.

Las fuentes de datos consideradas son PROSE con todos los registros que esta considera, Twitter, con todos sus registros, y por último Twitter Aseg. que son los vehículos denunciados por Twitter que cuentan con seguros particular.

Participación de robo según tipo de vehículo:

A continuación se presenta el gráfico de porcentaje de robos por tipo de vehículos, en donde se presentan las frecuencias que tienen cada una de las fuentes de datos.

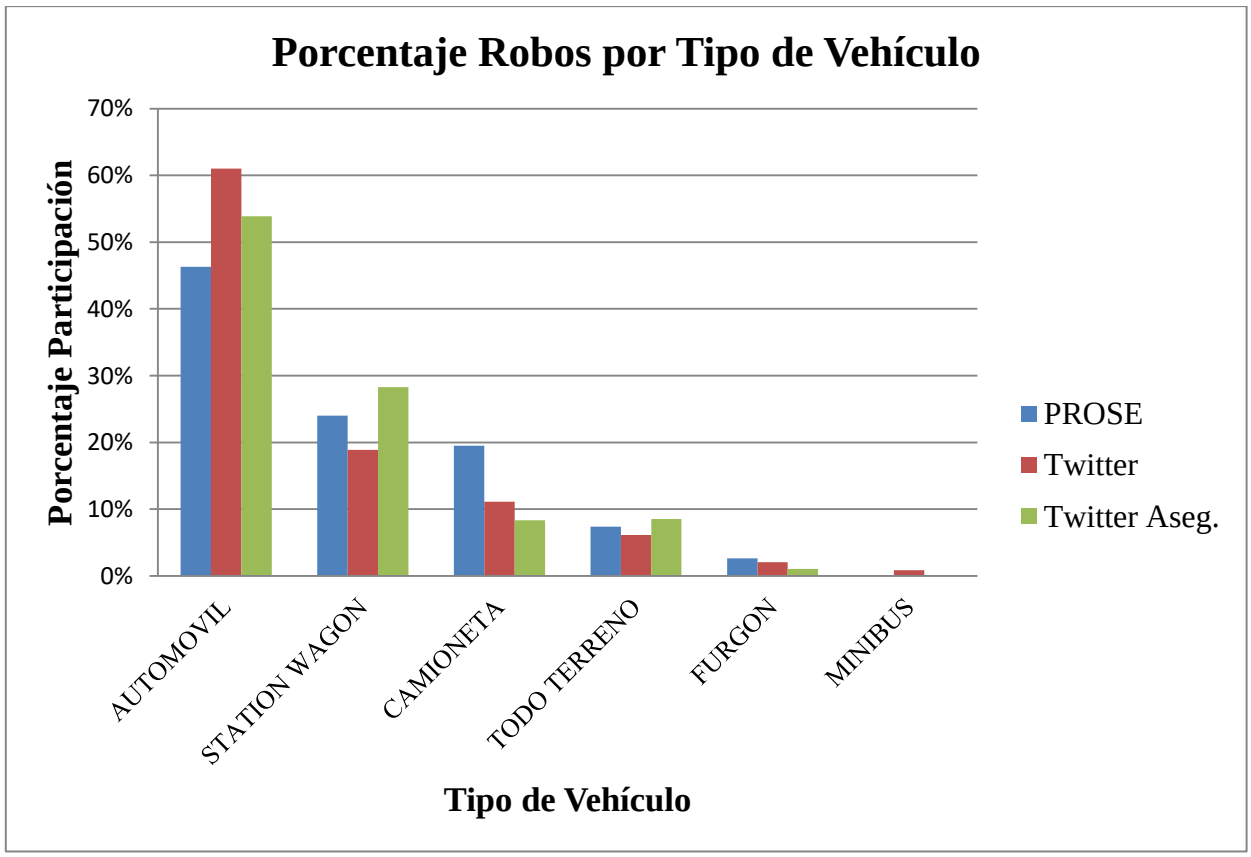

Figura 3.3: Gráfico Porcentaje de robos por tipo de vehículo.

En todas las fuentes de datos la distribución por categoría es similar, “AUTOMÓVIL” es la categoría de tipo de vehículo con mayor tasa de robos, y por otro lado la categoría “MINIBUS” es aquella que tiene menor tasa de robos entre 0% y 1%.

Participación de robo según grupo de tasación:

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

Recordar que se definieron 4 grupos de tasación acordes a los cuartiles de frecuencia en la base de datos de PROSE, y con esta definición de los cuartiles en la tasación se usó para las demás fuentes de datos para asignar un grupo de tasación.

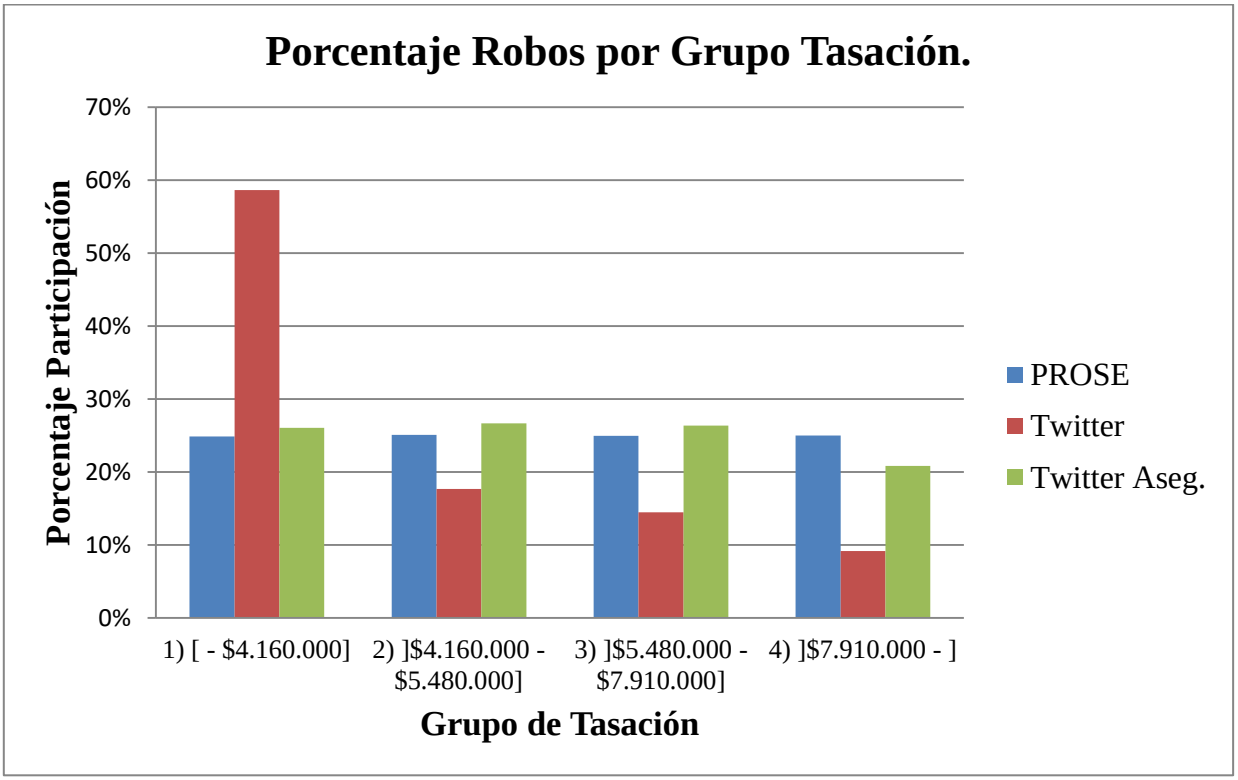

Figura 3.4: Gráfico Porcentaje robos por grupo tasación.

En el primer grupo de tasación, los de menor valor, Twitter presenta una altísima participación, una de las justificaciones posibles, es que como se mostró en 1.2, en el gráfico de distribución del tiempo en navegación por edad, 61% del tiempo de navegación en Internet es empleado por usuarios cuya edad se encuentra en el rango de 15-34, lo cual puede ser asociado a un segmento de menor poder adquisitivo.

Participación de robo según modelo de vehículo:

A continuación se presenta el gráfico de porcentaje de robos por modelos de vehículos, en donde se presentan las frecuencias que tienen cada una de las fuentes de datos.

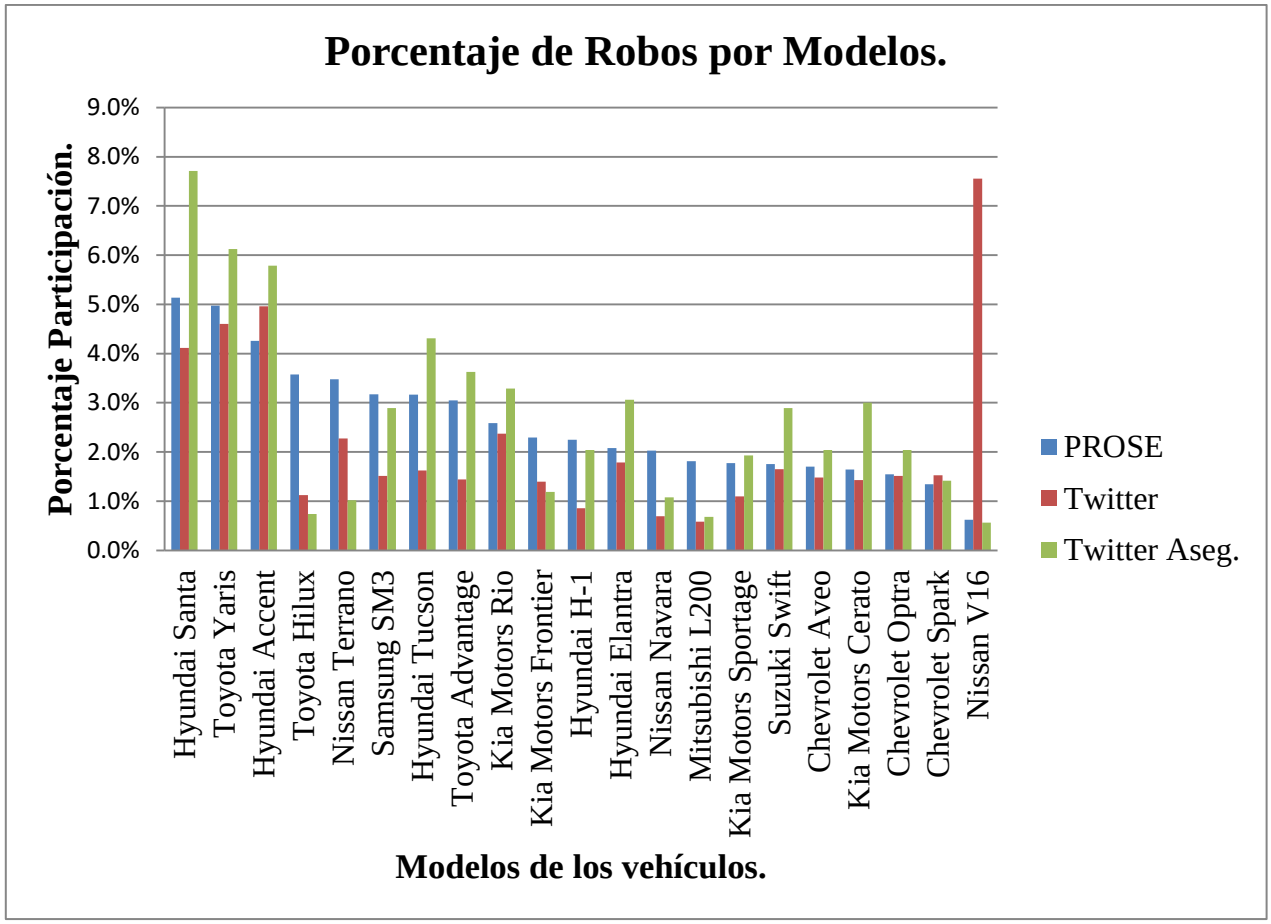

Figura 3.5: Gráfico Porcentaje de robos por modelos.

En términos generales se ve una relación directa entre las tasas de los modelos denunciados entre cada una de las fuentes de datos, siguen comportamientos similares, en mayor proporción entre PROSE y Twitter.

Existe un modelo que se escapa de toda relación con las otras fuentes de datos, es el último modelo visible, el vehículo modelo “Nissan V16”, en donde Twitter lo considera como el vehículo más denunciado.

La justificación es que uno de los vehículos más utilizado en Chile como Taxi es el modelo V16. En general los taxistas no aseguran sus vehículos ya que los valores de los seguros para los vehículos utilizados para transporte de pasajeros, ya sea la cuota mensual o el deducible, aumentan drásticamente en comparación a un vehículo de uso particular, principalmente por ser vehículos que están gran parte del tiempo del día siendo utilizados en las calles y por lo tanto están más expuestos a posibles accidentes o daños.

El modelo Nissan V16 es altamente robado, pero al ser usado principalmente como vehículo Taxi, es que en las fuentes de datos de seguros particulares no aparece con una tasa frecuente. En particular en PROSE o en Twitter Aseg. no aparece con tasas altas de frecuencias de robo este modelo.

Test de igualdad de proporciones:

Una de las inquietudes o una de las preguntas 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.

La pregunta nace ya que la red social permite informar a otros usuarios sobre eventos sucedidos, en este caso el robo de un vehículo, y esta información se puede empezar a masificar con alta rapidez por el uso de la aplicación.

Por otra parte se podría pensar que mientas más personas estén informadas del robo del vehículo y halla más observadores en las calles que pueden notificar el avistamiento del vehículo, será más posible encontrar el vehículo.

Para analizar las tasas de proporciones es que se realiza un Test de Igualdad de Proporciones sobre las proporciones de vehículos robados que fueron encontrados, comparando entre aquellos que solo fueron denunciados por las instituciones formales y aquellos que además de hacerlo por esta vía lo realizaron también por Twitter.

Es posible realizar dos tipos de test, el de una cola y el de dos colas, la diferencia está en la región de rechazo para un nivel de significancia. Es decir, la región de rechazo de la hipótesis nula, y aceptación de la hipótesis alternativa para el nivel de significancia escogido.

Cuando la hipótesis alternativa es del estilo “>” o “<” es que la región de rechazo queda definida por un extremo y por ende se habla de un test de una cola, en caso contrario en que la hipótesis alternativa se defina por “≠”, entonces la región de rechazo estaría definida por dos extremos y por ende un test de dos colas.

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 Si fueron denunciados por Twitter y fueron encontrados.*

$$\begin{array}{ll} Hipótesis\ nula & H_0\colon\ p_1 = p_2. \\ Hipótesis\ alternativa & H_1\colon\ p_1 < p_2. \end{array}$$

Como hipótesis alternativa se expresa que la proporción de hallazgos de aquellos vehículos robados que fueron denunciados por twitter es mayor que aquellos no lo fueron. Esta hipótesis alternativa genera un test de 1 cola al ser una hipótesis con desigualdad “<”.

Sea:

$$\hat{p}_i : \frac{x_i}{n_i}\ , la\ proporción\ de\ vehículos\ en\ la\ muestra\ i\ que\ fueron\ hallados.$$

$$SE(\hat{p}_i) = \sqrt{\frac{p_i(1-p_i)}{n_i}}\ , error\ estandar\ del\ estimador.$$

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}\ , estadístico\ del\ test.$$

$$\hat{p} : \frac{x_1 + x_2}{n_1 + n_2}\ , es\ un\ estimador\ de\ p.$$

Bajo la hipótesis nula, $\hat{p}$ es el estimador más preciso para estimar $p_1 = p_2 = p$.

Distribución nula: Bajo H_0, y para n_1 y n_2 suficientemente grande, el estadístico del test sigue una distribución z~N(0,1).

No es posible definir de manera arbitraria cuando el n_i es suficientemente grande, sin embargo en diversas investigaciones utilizan las fórmula de Cochran para poder estimar el número suficiente.

Fórmula de Cochran para tamaño de muestra [25]:

Para poblaciones de estudio que son grandes, Cochran (1963:75) desarrolló la siguiente ecuación para una muestra representativa de proporciones.

$$n_0 = \frac{z^2 pq}{e^2}$$

n_0 es el tamaño de la muestra, z^2 es el valor de la coordenada de la abscisa en la curva de la distribución normal para un área α en la cola (1- α es igual al nivel de confianza deseado, por ejemplo 95%), e es el nivel deseado de precisión, p es la proporción estimada de un atributo que está presente en la población de estudio, y q=1-p.

En el caso de esta investigación:

$z^2 = (1{,}96)^2$,lo cual corresponde a un intervalo de confianza del 95%.

$e^2 = (0{,}05)^2$,lo cual corresponde a un margen de error del 5%.

$p = 0{,}5$,es la proporción de hallazgos de vehículos, se asume 0,05 para considerar la máxima varianza en la muestra, sería el peor caso posible.

$q = 0{,}5$, es la proporción de los no hallados, que por ser q=1-p, es el mismo valor que p.

Por lo tanto la fórmula entrega el siguiente valor:

$$384{,}16 = \frac{(1{,}96)^2 0{,}5\ 0{,}5}{0{,}05^2}$$

Es decir se requiere un tamaño de muestra de 384 registros de robos de vehículos para considerar un tamaño de muestra suficientemente grande para que el estadístico del test de proporciones distribuya normal y se pueda realizar el test.

Como tanto la base de datos de PROSE como la base de datos de Twitter, superan los 385 registros, se asume normalidad del estimador con los datos estudiados.

Resultados del test de proporciones:

Para analizar las tasas de hallazgo se consideraron en la base de datos de Twitter a aquellos vehículos que están asegurados, es decir que las patentes coinciden en la base de datos de PROSE, los cuales son 1763 vehículos (24% de los datos totales).

Las proporciones de hallazgos de ambas bases de datos son:

PROSE: tasa de hallazgo del 59%.

Twitter: tasa de hallazgo del 70%.

Al realizar el test de proporciones a un 99% de confianza asumiendo distribución normal del estimador se obtiene como resultado un p-valor $< 2{,}2 * 10^{-16}$.

El p-valor es el valor mínimo bajo el cual es posible rechazar la hipótesis nula, en este caso el p-valor es menor a 0,01 (99% significancia), por lo tanto se rechaza la hipótesis nula de igualdad de proporciones y se acepta la hipótesis alternativa H_1: $p_1 < p_2$. Concluyendo que la diferencia de 11 puntos porcentuales entre ambas fuentes de datos si es significativa.

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 realiza un gráfico de 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.

Porcentaje de Hallazgos por Grupo de Tasación:

Para realizar esta comparación se consideran los registro de Twitter que también están presentes en PROSE, es decir, las tasas de hallazgos mostradas a continuación corresponden solo a vehículos con seguro particular.

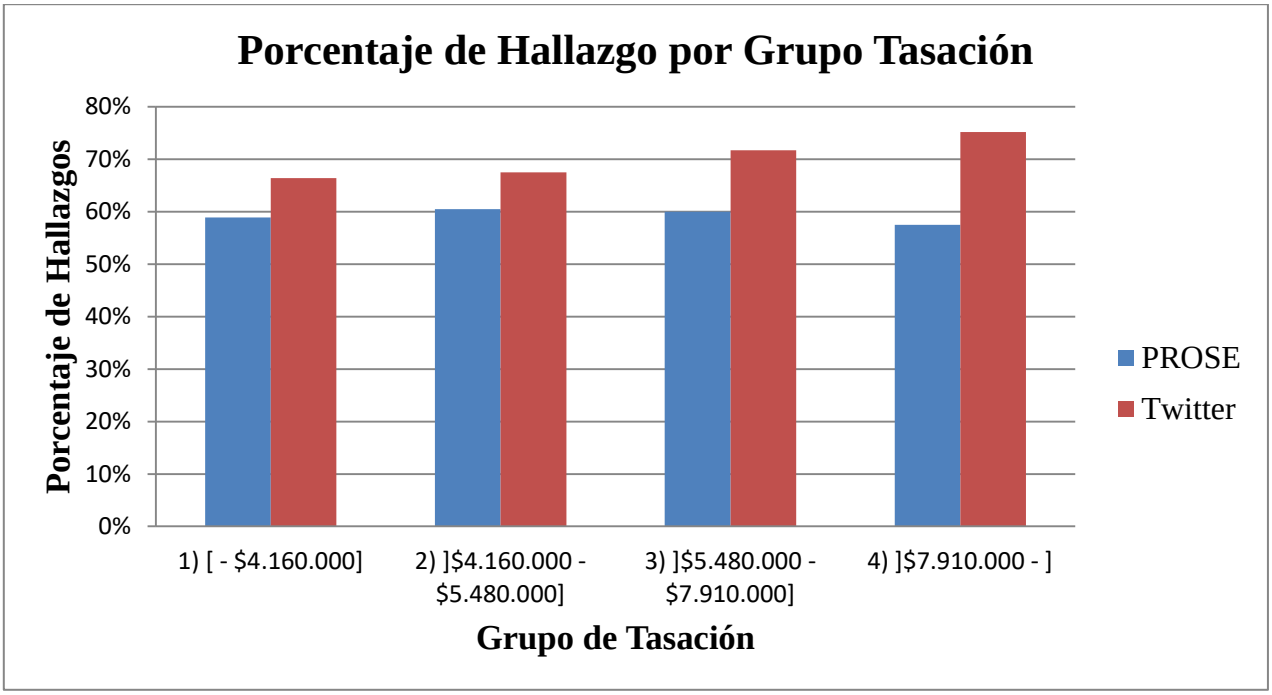

Figura 3.6: Gráfico Porcentaje de hallazgo por grupo tasación.

En el grafico se ve que todos los grupos de tasación aumentan su porcentaje de hallazgos, de hecho el grupo que más aumenta la diferencia porcentual es el grupo 4, el cual tiene menor tasa de recuperación.

Con el gráfico se puede demostrar que las diferencias porcentuales entre ambas fuentes de datos no se deben a un sesgo de los vehículos denunciados, al menos no en cuanto al valor de la tasación del vehículo.

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.

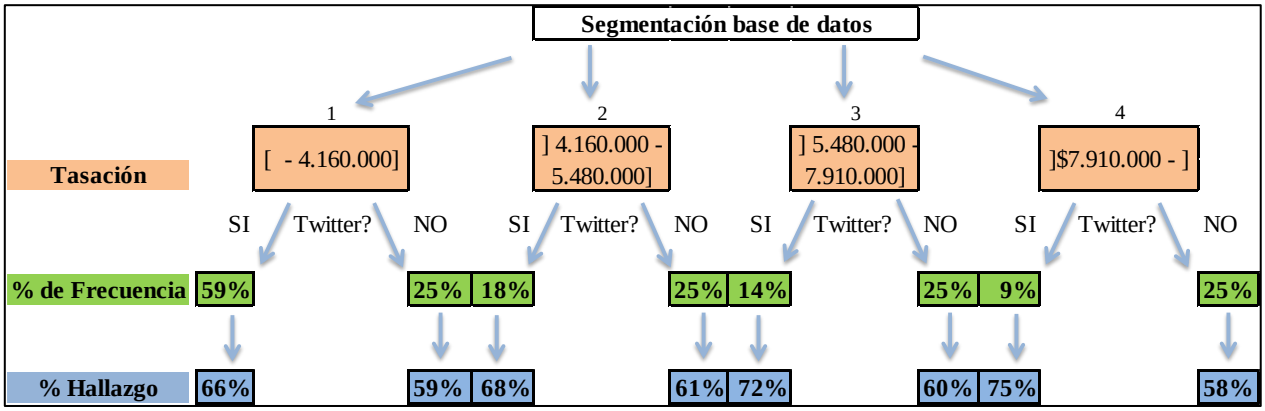

Figura 3.7 Resumen Árbol de análisis según grupo de tasación.

La Figura 3.7 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 PROSE de tal manera de que cada grupo representara un cuarto de la base de datos.

Análisis de la variación porcentual del Tipo de Vehículo robado según tipo de vehículo:

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 quizás 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 PROSE y posteriormente la fuente de datos de Twitter para observar si ambas presentan los mismos resultados.

A continuación se presentan los resultados para la fuente de datos de PROSE:

	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%**

Tabla 3.3: Variación tasa de robo de vehículo según tipo, en fuente de datos PROSE.

En los resultados de PROSE se puede visualizar que dos tipos de vehículos presentan variaciones importantes, el tipo de vehículo “Automóvil” y el tipo “Camioneta”

De hecho se compensan las variaciones, el fin de semana, se puede observar que la tasa de robo del tipo “Automóvil” aumenta, sin embargo esto hace disminuir la tasa del 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 PROSE, se mostrarán los resultados de Twitter para poder observar si estos muestran los mismos.

	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 3.4 Variación tasa de robo de vehículo según tipo, en fuente de datos Twitter.

Los resultados de Twitter son similares, solo 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.

3.3.6 Interpretación de patrones.

En este paso final se interpretan los patrones y relaciones encontrados en el proceso de minería de datos. Se han aplicado diferentes técnicas estadísticas 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.

Reconstrucción del hecho:

Las fases del hecho son Robo del Vehículo – Envío del Tweet – Denuncia en Carabineros de Chile – Validación PROSE – Hallazgo del Vehículo.

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 realizar la denuncia 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 cuando el delito es realizado.

Envío del Tweet: Es la hora exacta en la que el Tweet 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 PROSE: Es la fecha y hora en que PROSE 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 PROSE, quien las valida y almacena.

Hallazgo del Vehículo: Corresponde a la fecha y hora en la que el vehículo denunciado por robo fue hallado. Tal como se mostró anteriormente, no todos los vehículos son encontrados, por lo tanto para este cálculo se consideraron aquellos que efectivamente fueron encontrados.

Para medir los tiempos entre las fases se calculó la mediana de cada una, se prefirió utilizar esta medida en vez del promedio simple, ya que hay valores extremos, conocidos como "outliers", los cuales provocan que el promedio indique un valor que se aleja de la generalidad de los datos, y en este caso se pretende entender el orden de los hechos en su generalidad. Un ejemplo de los datos que desproporcionan la medida es el caso de un vehículo que fue encontrado 3 años después de la fecha en que fue robado, evidentemente este dato incrementará considerablemente el promedio en el tiempo transcurrido para hallar el vehículo. El cálculo de la mediana, permite evitar estos problemas al ser una medida que se basa en el ordenamiento de todos los tiempos y luego

considerar el que se sitúa en el medio del listado, de esta forma considera como tiempo general a aquel que se ubica en el centro.

La reconstrucción del hecho indica el siguiente orden y sus respectivos tiempos entre las fases:

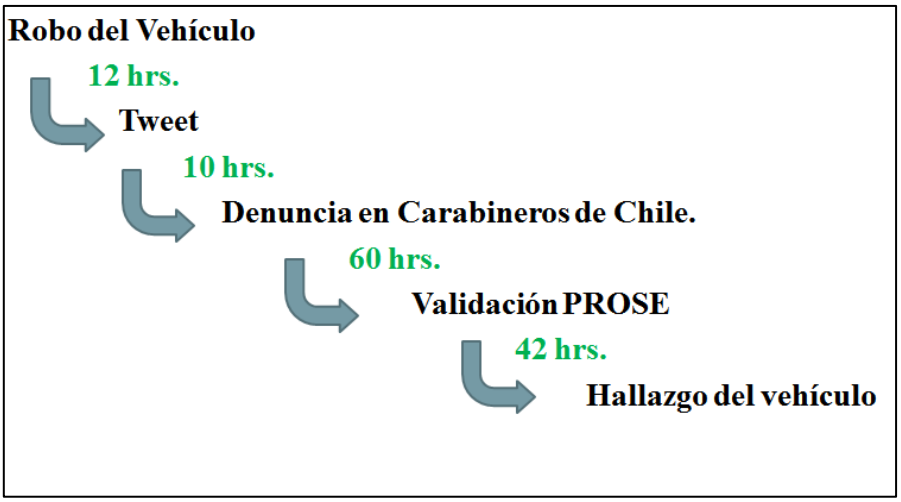

Figura 3.8: Reconstrucción orden de los hechos en un robo de vehículo.

Es decir el orden de los hechos es:

1. Robo del vehículo.
2. Envío del Tweet.
3. Denuncia en Carabineros de Chile.
4. Validación PROSE.
5. Hallazgo

El resultado no era evidente, existía la pregunta sobre si el envío del Tweet era realizado antes de la denuncia en Carabineros de Chile, y ahora se acaba de corroborar. Esto expresa el valor que presenta Twitter, ya que permite obtener información de manera anticipada, de hecho la mediana indica que el Tweet se genera con 10 hrs. de anticipación frente a la denuncia formal, lo cual es bastante tiempo.

El hecho de que el Tweet sea el primer evento generado luego del robo del vehículo hace que Twitter sea visto como una fuente de información valiosa por las características temporales de cómo se construye el hecho y se va obteniendo la información.

Además hay un largo periodo de tiempo transcurrido entre que el robo es realizado y PROSE 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 PROSE obtiene la información, en muchos casos la obtención de esos datos se aproxima a las 2 semanas (ver anexo 1), 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 PROSE, en la mayoría de los casos es cercana a los 5 días. Esta información

puede ser utilizada para focalizar los esfuerzos en hallar los vehículos dentro de esos 5 días principalmente.

Interpretación de los resultados presentados en la minería de datos:

1) Inspección de las tasas de frecuencia

Se realizó una inspección visual y luego se calculó la correlación de las tasas de denuncias mensuales entre la base de datos de PROSE y Twitter en el periodo 2012 – 2016.

Visualmente se observa un comportamiento de correlación entre ambas fuentes de datos, lo cual se acentuaba más en el periodo 2015 – 2016.

Al calcular la correlación demostró lo que visualmente se veía, que hay correlación positiva entre ambas fuentes de datos, y mayor en el último periodo.

La visualización ya permitía observar lo que el factor de correlación mostró finamente, por lo que es prudente considerar estos gráficos en el proyecto del observatorio si se quiere ir monitoreando en todo momento como están cambiando las tasas de robos en ellos se observan claramente los crecimientos o los decrecimientos que van surgiendo.

Al tener una alta correlación ambas fuentes de datos, permite considerar a Twitter como una buena fuente de información para esta medida, teniendo en cuenta que la obtención de estos datos es anticipada.

2) Sesgos en las bases de datos.

Tipo de vehículo y tasación:

Los resultados mostraron que twitter sigue una correlación directa con los datos de PROSE sin embargo sobre muestrea el tipo "Automóvil" en un 15% aproximadamente, lo cual disminuye la participación de los otros tipos.

Además un sesgo relacionado se mostró en el gráfico de la tasación de los vehículos, en donde el primer grupo ([- $4.160.000]) está sobre representado, lo cual calza con el tipo “Automóvil” ya que este tipo de vehículo es el que está más presente en el primer grupo de tasación.

Es muy importante que se tengan en consideración estos dos sesgos, principalmente el de la tasación de los vehículos ya que podría afectar considerablemente un modelo predictivo, estar consciente de que Twitter como fuente de datos muestra preferentemente el grupo de menor valor, y por consecuencia los de mayor valor, son considerablemente menos denunciados en comparación con las que recopila PROSE.

Modelos de vehículos:

En general presentan una correlación alta, los modelos más robados presentan tasas similares en ambas fuentes de datos, el caso en donde estos se desproporciona es en el modelo "Nissan V16", la razón de este comportamiento de debe a que este modelo de vehículo es principalmente utilizado para el transporte de público de pasajeros menor, "taxi", y estos vehículos suelen no

asegurarse porque los seguros particulares para los vehículos que son utilizados para estos fines son más elevados.

Lo importante de este sesgo es estar al tanto de los modelos que se están utilizando para uso de transporte de pasajeros ya que estos presentarán considerablemente mayor porcentaje de participación en la fuente de datos de Twitter en comparación con la de PROSE. Por ende si se vuelve a observar que un modelo de vehículo se está usando principalmente para el transporte de vehículos este presentará tasas bajísimas en los datos de vehículos asegurados en comparación a la fuente de datos de Twitter.

Impacto del envío de Tweet en tasa de Hallazgo:

Se demostró que los vehículos que son denunciados en Twitter a través de un Tweet, tienen mayor tasa de recuperación que aquellos que no lo son, independiente del grupo de tasación al cual corresponda el vehículo, la tasa de hallazgo aumenta significativamente, de hecho el grupo de mayor valor (]$7.910.000 -]), el cual es el que tiene menor participación de denuncias en Twitter, es el que presenta mayor aumento en la tasa de recuperación.

Con esta información se podría realizar una campaña informativa o de recomendaciones para motivar su uso para fines de hallazgo de vehículos robados.

Análisis diario sobre la variación porcentual del tipo de vehículo robado:

Al realizar el análisis sobre como varían las tasas de robo de los tipos de vehículos según el día de la semana, los resultados mostraron que hay un efecto notorio en los fin de semana. Durante el fin de semana, la categoría "Automóviles" aumenta, por el contrario la categoría "Camioneta" se ve disminuida en los mismos días.

Lo importante del comportamiento es que se refleja en ambas fuentes de datos, tanto en PROSE como en Twitter. Por ende se puede ir monitoreando el comportamiento día a día de manera anticipada con Twitter.

Respecto de la variación de los otros tipos de vehículos, en ambas fuentes de datos se comportaron de manera similar, sin mayores variaciones durante los días de la semana. Por lo tanto twitter muestra el mismo comportamiento diario con respecto a los datos de PROSE, respecto de la variación en los tipos de vehículos.

Capítulo 4: Conclusiones y limitaciones

4.1 Conclusiones

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

A pesar de que en investigaciones anteriores (ver 2.2) tuvieron problemas para catalogar los Tweets en la categoría "Robo de Vehículo" y por consiguiente en medir sus frecuencias de denuncias, en esta tesis se lograron extraer 18.112 registros relacionados al robo de vehículos de los cuales se verificó que el 66% correspondían a denuncias de robos de vehículos con la respectiva patente incluida.

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 PROSE, 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 factor 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, en donde más del 60% del tiempo navegado está distribuido en aquellos menores a 35 años (ver sección 1.2), 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. Hay un caso excepcional en el modelo Nissan V16 el cual es utilizado frecuentemente para transporte de pasajeros, "taxi", y estos no suelen ser asegurados, por ende no aparecen en los registros de PROSE con la misma frecuencia.

Lo relevante del hallazgo es que para los modelos más robados, Twitter presenta tasas muy similares a las denuncias realizadas en PROSE, 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 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 PROSE 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, justificando su utilización en el proyecto en el cual se enmarca esta investigación.

El hecho de que el Tweets se origine como el primer evento luego de la sustracción del vehículo puede ser aprovechado por distintas entidades como PROSE, 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.

El proyecto en el cual se enmarca esta investigación desarrollará un modelo predictivo del robo de vehículos, y en esta tesis se mostró información relevante sobre las relaciones existentes entre PROSE y Twitter.

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.

La metodología utilizada para realizar las coincidencias en los modelos, basada en "Levenshtein distance" es un entregable para PROSE ya que para ellos es una problemática a resolver al momento de realizar las mediciones, ya que como se mencionó, la escritura de los modelos no están homologados en los registros vehiculares, dificultando el cálculo de las estadísticas relacionados a estas variables.

Los objetivos propuestos al inicio de esta tesis fueron cumplidos exitosamente.

El primer objetivo sobre identificar las metodologías aplicadas y los resultados obtenidos en investigaciones relacionadas con los delitos utilizando Twitter como fuente de información fue cumplido, en el capítulo 2.2 se muestran estos resultados, además ayudaron a guiar el desarrollo estadístico de esta tesis.

El segundo objetivo sobre investigar y aplicar métodos de extracción de Tweets fue cumplido, se mostraron las diferentes opciones para extraer Tweets, y se desarrolló un script propio, el cual permitió obtener la base de datos de Twitter con 18.112 registros.

El tercer objetivo sobre aplicar herramientas de minería de texto sobre la base de datos que contiene los Tweets, fue realizado con éxito, aplicando diferentes metodologías se extrajeron las patentes de los vehículos denunciados en el contenido del Tweet, y se aplicó "Levenshtein distance" en variables de tipo no numéricas, entre otras técnicas de aplicadas.

El cuarto objetivo dirigido a evidenciar los sesgos encontrados en Twitter al compararlo con PROSE fue cumplido en la sección 3.2.5 de minería de datos, en donde se mostraron diferentes sesgos que tiene la red social respecto de las características de los vehículos denunciados.

La hipótesis de investigación fue corroborada, al mostrar las relaciones existentes entra ambas fuentes de datos, un factor de correlación de 0,73 en las tasas de denuncias, y similitud en los patrones de comportamientos.

4.2 Limitaciones

Algunas de las limitaciones experimentadas en el desarrollo de esta investigación está el hecho de que al momento de extraer los Tweets, se aplicó una metodología que exige definir palabras filtros, se definieron 3 conceptos: *robo patente OR robado patente OR robaron patente.* El haber definido estos conceptos limitó la extracción de Tweets, se podría realizar una extracción con más registros y observar los resultados, sin embargo se espera obtener conclusiones similares.

Otra limitación de esta investigación está en la homologación de los modelos, en donde se debió aplicar Levenshtein distance, para realizar la coincidencia de variables no numéricas, si bien al observar los resultados se apreció un comportamiento correcto, este no es perfecto, y podría homologar modelos incorrectos en casos en que el modelo esté abreviado o escrito de manera poco similar a su modelo correcto. Las repercusiones de esto recaen en la tasación de los vehículos la cual se realiza con estos modelos homologados, por ende por error de arrastre podría significar una mala tasación, y errores en las estadísticas.

Finalmente una de las limitaciones que enfrenta esta investigación es sobre los vehículos que fueron considerados en el análisis, se consideraron solo los vehículos livianos, debido a que los vehículos pesados contemplaban una gran cantidad de errores en los datos registrados en la variable que corresponde a los modelos de los vehículos, lo que provocaba que sus datos no fueran fidedignos, o implicaban una mala homologación de estos. Además estos vehículos pueden variar su valorización de $5.000.000 a $60.000.000 al aplicar una mala homologación, por lo tanto podrían implicar grandes errores en el análisis. Es por esto que, considerando que este tipo de vehículo representaba el 9% del total de registros, se decidió por focalizar el estudio en los vehículos livianos.

Bibliografía

[1] Subsecretaría de Telecomunicaciones, «Accesos a Internet llegan a 13,1 millones y uso de smartphones sigue en alza según estadísticas de telecomunicaciones,» [En línea]. Available: http://www.subtel.gob.cl/accesos-a-internet-llegan-a-131-millones-y-uso-de-smartphones-sigue-en-alza/. [Último acceso: Abril 2017].

[2] comScore, «Futuro Digital, Chile,» 2014.

[3] Telefonica, «Informe Big Data de Movistar Chile: CHILENOS PASAN MÁS DE 4 HORAS AL DÍA OCUPANDO SUS SMARTPHONES,» [En línea]. Available: Chile http://www.telefonicachile.cl/informe-big-data-de-movistar-chile-chilenos-pasan-mas-de-4-horas-al-dia-ocupando-su-smartphone/. [Último acceso: Abril 2017].

[4] Telefónica , «Big Data,» 2016. [En línea]. Available: http://www.telefonicachile.cl/wp-content/uploads/2016/12/Informe-Big-Data_20161.pdf. [Último acceso: Mayo 2017].

[5] Carabineros de Chile, «Algunas definiciones,» [En línea]. Available: Delitos de mayor connotación social http://dac.carabineros.cl/datos.php. [Último acceso: Diciembre 2016].

[6] ENUSC, «Presentación de Resultado XII ENCUESTA NACIONAL URBANA DE SEGURIDAD CIUDADANA,» 2015. [En línea]. Available: http://www.ine.cl/docs/default-source/sociales/seguridad-ciudadana/2015/enusc_xii_resultados.pdf?sfvrsn=9. [Último acceso: Diciembre 2016].

[7] Subsecretaría de Prevención del Delito, «Delitos de Mayor Connotación Social - Series de datos 2001 - 2016,» [En línea]. Available: http://www.seguridadpublica.gov.cl/estadisticas/tasa-de-denuncias-y-detenciones/delitos-de-mayor-connotacion-social-series-de-datos-2001-2015/. [Último acceso: Abril 2017].

[8] «INE ANUARIOS PARQUE DE VEHÍCULOS EN CIRCULACIÓN 2015,» [En línea]. Available: http://www.ine.cl/canales/chile_estadistico/estadisticas_economicas/transporte_y_comunicaciones/parquevehiculos.php# . [Último acceso: Abril 2017].

[9] U. Fayyad, G. Piatetsky-Schapiro y P. Smyth, «From Data Mining to Knowledge Discovery in Databases,» *AI Magazine* , vol. 17, nº 3, pp. 37-54, 1996.

[10] Twitter, «Blog Twitter,» [En línea]. Available: https://twitter.com/twittersupport/status/555076845293432834. [Último acceso: Marzo 2017].

[11] Twitter, [En línea]. Available: https://twitter.com/Victor_Curico/status/853382736236752896. [Último acceso: 1 Agosto 2017].

[12] M. Coletto, c. Lucchese, S. Orlando y R. Perego, «Electoral predictions with twitter: a mahine-learning approach,» de *Proceedings of the IIR 2015*, Cagliari, Italia, 2015.

[13] A. Culotta, «Towards detecting influenza epidemics by analyzing Twitter messages,» de *SOMA'10 Proceedings of the First Workshop on Social Media Analytics*, Washington D.C, 2010.

[14] R. Todd-Bennett, «Identifying Crime Hotspots,» 2015.

[15] N. Malleson y M. A. Andresen, «The impact of using social media data in crime rate calculations: shifting hot spots and shanging spatial patterns,» *Cartography and Geographic Information Science*, vol. 42, nº 2, pp. 112-121, 2015.

[16] M. Gerber, «Predicting Crime using Twitter and Kernel Density Estimation,» *Decision Support Systems*, vol. 61, pp. 115-125, 2014.

[17] Twitter, «Twitter Ads API,» [En línea]. Available: https://dev.twitter.com/ads. [Último acceso: Marzo 2017].

[18] Twitter, «REST APIs,» [En línea]. Available: https://dev.twitter.com/rest/public. [Último acceso: Marzo 2017].

[19] Twitter, «Streaming APIs,» [En línea]. Available: https://dev.twitter.com/streaming/overview. [Último acceso: Marzo 2017].

[20] Twitter, «robot.txt,» [En línea]. Available: https://twitter.com/robots.txt. [Último acceso: Marzo 2017].

[21] Twitter, «Search Twitter,» [En línea]. Available: https://twitter.com/search?q=%23. [Último acceso: Marzo 2017].

[22] D. Padula y L. Debera, «TÉCNICAS DE IMPUTACIÓN, UNA APLICACIÓN PARA MEDIR EL INGRESO,» de *Décimo Congreso Latinoamericano de Sociedades de Estadística.*, Córdoba, 2012.

[23] R. Little y D. Rubin, de *Statistical Analysis witch Missing Data*, Segunda ed., New Jersey, Wiley, 2002, p. 381.

[24] Servicio de Impuestos Internos, «Sii Tasación fiscal de vehículos,» [En línea]. Available: http://www.sii.cl/pagina/actualizada/noticias/tasacion_vehiculos.htm . [Último acceso:

Abril 2017].

[25] 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.

[26] «The R Project for Statistical Computing,» [En línea]. Available: https://www.r-project.org/ . [Último acceso: Mayo 2017].

[27] «The R environment,» [En línea]. Available: https://www.r-project.org/about.html. [Último acceso: Mayo 2017].

[28] «RSelenium Introduction,» [En línea]. Available: https://cran.r-project.org/web/packages/RSelenium/vignettes/RSelenium-basics.html. [Último acceso: Mayo 2017].

Anexos:

A Entorno de Trabajo

Para el desarrollo de esta tesis se utilizó el siguiente entorno de trabajo.

Sistema operativo:

Microsoft Windows 7 Professional.

6 GB de memoria RAM.

500 GB de disco duro.

Procesador Intel i5-2410M.

Software y lenguaje de programación:

Para realizar todo el proceso KDD se utiliza el software estadístico R. R es un software gratuito para realizar computación estadística y gráficos [26].

R es un proyecto colaborativo, en donde desarrolladores pueden añadir funcionalidades definiendo nuevas funciones las cuales son almacenadas en librerías [27].

El acceso a las librerías generadas por otros usuarios es gratuito y se pueden descargar para poder ser utilizadas con fines personales.

Librerías utilizadas:

RSelenium: Esta librería tiene como objetivo facilitar la conexión con un servidor Selenium dentro del software R. Selenium es un proyecto focalizado en automatizar navegadores web [28].

Con el uso de RSelenium se logra manipular un navegador web a través del software R, permitiendo extraer contenido HTML de páginas web.

Readxl: Librería que permite leer archivos de Excel, en formato ".xlsx".

Stringdist: Librería que permite aplicar diferentes métricas de distancia, entre ellas Levenshtein distance, la cual fue utilizada en esta tesis.

B Comparación de las tasas de emisión de denuncias y las tasas de obtención de esa información en PROSE, medido diariamente por 2 meses (Noviembre 2016 – Diciembre 2016).