UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

ESTUDIO DE PATRONES Y RELACIONES MEDIANTE UN ANÁLISIS CUANTITATIVO ENTRE MÉTRICAS DE EVENTOS SÍSMICOS CON DATOS EXTRAÍDOS DE TWITTER

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN CIENCIAS, MENCIÓN COMPUTACIÓN

MEMORIA PARA OPTAR AL TITULO DE INGENIERO CIVIL EN COMPUTACIÓN

JUAN IGNACIO VALDERRAMA LORCA

PROFESOR GUÍA:
BARBARA POBLETE LABRA
MARCELO MENDOZA ROCHA

MIEMBROS DE LA COMISIÓN:
AIDAN HOGAN
GONZALO NAVARRO BADINO
HANS LÖBEL DIAZ

SANTIAGO DE CHILE
2019

RESUMEN

Motivación. Twitter es una red social que permite a las personas la interacción con otras y la difusión de información del mundo real. Por otro lado, en los desastres naturales, surge la necesidad de una rápida y confiable recopilación de lo sucedido y sus consecuencias. Por lo anterior, Twitter es considerado una posible gran fuente de información para eventos físicos significativos, en particular, en desastres naturales como los sismos.

Objetivo. Dado el contexto, el objetivo de este trabajo es evidenciar la correspondencia medible entre la intensidad de un evento sísmico y la visibilidad del evento en Twitter. Se estudiará si esta correspondencia sea traducible en un modelo predictivo de intensidad de sismos.

Contribución. El trabajo es un puente entre los eventos sísmicos y Twitter. Permite mostrar las características de Twitter que son relevantes en un evento sísmico, y a su vez, establece una metodología para identificar correlación entre sismos y Twitter. Finalmente, plantea un método que logra la correcta relación entre eventos sísmicos y Twitter. Este trabajo es pionero en el uso de las redes sociales para la estimación de los daños provocados por un sismo.

Metodología. Se propone y aplica una metodología que considera desde la extracción de los datos hasta la generación de modelos y comparación de resultados. Primeramente, se extraen los sismos de un intervalo definido de tiempo, y a su vez, los mensajes asociados a dicho evento sísmico. Luego ubicamos geográficamente los mensajes y generamos las características que serán relacionadas con los sismos. Posteriormente, se identifican las ubicaciones que percibieron el sismo y sobre estas comunas, se crea un modelo para estimar la intensidad de Mercalli. Finalmente, mejoramos el modelo utilizando la dimensión espacial y comparamos los resultados.

Valor. El valor de este trabajo es ir más allá en las investigaciones que relacionan sismos con redes sociales, al incluir un mayor grado de profundización, al predecir la intensidad de los sismos a nivel de comunas y en la predicción restringida al uso exclusivo de información de Twitter.

Agradecimientos

Comienzo diciendo que quedaré corto agradeciendo a todas las personas que estuvieron en este proceso de mi carrera universitaria. A lo largo de la carrera he conocido muchas personas que han contribuido en llegar a este día, tanto en el ámbito académico como relaciones personales. Primero agradezco a mis amigos de sección y los que conocí a lo largo de plan común. Menciono en forma especial a Alonso Cubillos, Natalia Cheuquenao y Rodrigo Plaza, nos conocimos desde el primer día de universidad, y han sido un pilar en mi vida universitaria, gracias por la amistad que ha trascendido el ámbito universitario. Imposible olvidar a toda la gente de la salita del DCC, que siempre con su buenas vibras ayudaban en el día a día. Especialmente a Felipe Rodriguez, Luis Martinez, Caterina Muñoz y Milenko Tomic, amigos quienes pasamos gran parte de la especialidad juntos. Gracias por hacer más agradables esos momentos de estrés.

Gracias a mis profesores guías, Barbara y Marcelo. Desde la primera vez que me acerque a la oficina de Barbara buscando tema de tesis, estuvo dispuesta a ayudarme. Ambos fueron que fue un apoyo muy importante en mi trabajo, por su disposición, consejos y el aporte de ideas que me entregaron para lograr mi trabajo.

A mi familia, por apoyarme desde el primer segundo, que decidí embarcarme en esta carrera, y en cada una de las decisiones que he tomado en mi vida, tanto aciertos como desaciertos. Sé que nunca podre retribuir todo lo que han hecho por mi y ese amor incondicionales que me entregan, simplemente los amo.

Hago una mención muy especial a mi pareja Paulina Leppe, por su apoyo incondicional en todo este proceso. Ella ha vivido conmigo los buenos y malos momentos de todo este proceso, y siempre ha estado ahí para apoyarme. Gracias por toda tu paciencia, el leer y escuchar todo mi trabajo, el ayudarme a redactar y sobre todo amarme. Soy feliz de vivir esta proceso junto a ti, te amo.

Después de este largo camino, me quedo con el aprendizaje entregado por cada persona que conocí en esta etapa de mi vida. Simplemente Gracias.

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Capítulo 1

Introducción

1.1. Motivación

Twitter es una red social que surgió en el año 2006. A lo largo del tiempo, ha logrado un uso masivo a nivel mundial. En el año 2010, M. Okazaki et al. [2], plantaron la idea de utilizar Twitter para la detección de eventos del mundo físico, basándose en la instantaneidad que presenta las publicaciones de sus usuarios. En investigaciones posteriores, T. Sakaki et al. en los años 2010 [3] y 2013 [4], presentaron esta idea enfocada en la detección de eventos sísmicos, donde desarrolla la recolección y procesamiento de datos de Twitter. Además incluyó un modelo para identificar en tiempo real los eventos sísmicos y un modelo para la estimación del epicentro del sismo. En estos trabajos se utiliza Twitter como una herramienta de investigación científica y surge el concepto de *sensor social*, que se basa en la actividad registrada por los usuarios de Twitter como fuente de información. Los usuarios al ser estimulados en el mundo físico responden o se expresan mediante los tweets[1], siendo este punto donde se genera la información.

En el año 2010 investigadores del U.S. Geological Survey (USGS) [5] publicaron un artículo que consistió en un estudio de los datos recolectados del sismo que ocurrió en California en el año 2009. Su investigación describe las ventajas y desventajas de utilizar Twitter como fuente de detección de sismos. En dicha investigación, los reportes obtenidos desde Twitter se recogieron en menor tiempo que el promedio de detección de los sismos usando los sensores sismológicos. Por otra parte, identificaron la dependencia del número de usuarios que se encuentran en la zona afectada como una limitación que presenta Twitter como fuente de información. La conclusión de la investigación es que las redes sociales como fuente de detección de sismos, no son un reemplazo a los sensores físicos, pero pueden ser un buen complemento, ya que destaca la rapidez de detección de eventos, permitiendo producir información adicional en base a la percepción de las personas.

En el año 2014, L. Burks y M. Miller [6] compararon diferentes modelos de regresión para la estimación de la intensidad de sismos en base a los datos de Twitter y reportes

[1]Tweet: Breve mensaje utilizado para comunicarse en Twitter. Revisar la sección 2.2

entregados por sismógrafos a lo largo de Japón. En esta investigación utilizaron información de geolocalización de los tweets y las coordenadas del sismo, llegando a la conclusión de que el modelo más efectivo es la regresión elastic net.

En el año 2016, Y. Kryvasheyeu et al. [7] realizan un estudio para poder estimar el daño causado por un desastre natural. Los datos que tomaron fueron los tweets durante el huracán Sandy ocurrido el 2012. En su investigación concluyeron la directa relación entre el número de tweets con el valor monetario de los daños.

1.1.1. Descripción del problema

En los inicios de las redes sociales, su uso como fuente de información confiable fue cuestionado, pero en la actualidad es importante su rol en la investigación. Debido a lo anterior, se han logrado crear distintas plataformas para la detección de eventos, en particular sismos, basadas en la información que entregan los usuarios de Twitter, y así recopilar rápidamente la percepción de las personas sobre estos eventos. Dentro de estas plataformas, se encuentran EARS [8] y ESA [9], implementadas en Italia y Australia, respectivamente. Ambas plataformas presentan buenos resultados en el ámbito de detección y alerta de sismos, siendo un apoyo en el monitoreo de estos eventos. En esta investigación se pretende ir un paso más adelante, generando un modelo que estime métricas sismológicas, como la intensidad del evento a partir de la información entregada por Twitter.

Por otro parte, los algoritmos propuestos para la estimación del epicentro [4] o de la intensidad [8], parecen correctas soluciones, pero presentan claras limitaciones que se desean superar en esta investigación. Queremos ser independientes de elementos de la red de sismógrafos, algo que limita a [4]. Para el caso de [8], solo estima la intensidad máxima del sismo, lo que se traduce en la estimación exclusivamente de una zona dentro del sismo, por lo que esperamos extender la estimación en distintas ubicaciones dentro del evento sísmico.

1.2. Objetivos

1.2.1. Objetivo general

El objetivo principal de la tesis es validar la correspondencia medible entre la intensidad de un evento sísmico y la visibilidad del evento en Twitter.

1.2.2. Objetivos específicos

1. Investigar la correlación que puede existir entre las métricas de los tweets que hablen de un evento sísmico con las magnitudes Richter y Mercalli que presenta dicho evento. Algunas métricas a relacionar con sismo son: cantidad de tweets, cantidad de palabras,

uso de signos de exclamación, entre otros.

2. Generar un modelo predictivo para la intensidad de eventos sísmicos a partir de Twitter, identificando las mejores métricas de los tweets y combinaciones de estas para obtener el modelo más preciso.

3. Extender el modelo considerando la dimensión geográfica dentro del sismo, y así mejorar los resultados de las métricas sísmicas.

1.3. Metodología

La metodología que busca cumplir los objetivos mencionados anteriormente se separa en tres etapas. La primera etapa es la definición y recolección de datos, enfocada en recopilar la información de eventos sísmicos y obtener los tweets de Twitter relacionados a estos eventos, utilizando como fuente el sistema de detección de eventos sísmicos del CSN[2]. Con esta información se busca establecer un conjunto de métricas que se utilizarán en la etapa siguiente. La segunda etapa, estimación de Mercalli, consiste en generar modelos predictivos en base a la información recolectada en la primera etapa. Esta etapa se descompone en predecir las comunas de interés y posteriormente, la predicción de Mercalli sobre estas comunas. La tercera etapa consiste en mejorar los distintos modelos creados en la etapa anterior incorporando la dimensión espacial del fenómeno.

A continuación se muestra el detalle de cada etapa.

1. Etapa (extracción de características)

    - Extraer los datos obtenidos por el sistema de detección de eventos sísmicos del CSN, que considera los Tweets del 2016 relacionados con eventos sísmicos, que se encuentran almacenados en una base de datos relacional.

    - Extraer la información del Centro Sismológico Nacional sobre los distintos sismos ocurridos en Chile.

    - Vincular los tweets correspondientes a cada evento sísmico.

    - Identificar la ubicación geográfica de los tweets.

    - Identificar métricas que se pueden extraer de los datos de Twitter para nutrir el futuro modelo que se creará.

2. Etapa (estimación de Mercalli)

    - Crear modelo para identificar las comunas que son percibidas en un sismo, denominados regiones de interés.

[2]CSN: Centro Sismológico Nacional

- Crear modelo para estimar magnitud Mercalli a partir de la regiones de interés.

3. Etapa (estimación espacial de Mercalli)

   - Aplicar sobre los resultados de la estimación de Mercalli un método que considera la dimensión espacial del fenómeno.

   - Comparar precisión entre modelos.

   - Identificar posibles razones frente a los resultados entregados por los modelos.

1.4. Estructura de este trabajo

Este trabajo esta organizado de la siguiente manera:

1. En el capítulo 2, definimos conceptos de Twitter y sismos. Además se discute acerca de herramientas para el análisis de datos como clasificadores y criterios de evaluación.

2. En el capítulo 3, describimos el trabajo relacionado de la tesis, donde mostramos las investigaciones realizadas con la red social Twitter, y trabajos previos con análisis de datos en desastres natural, en particular, sismos.

3. En el capítulo 4, describimos detalladamente la metodología empleada en la investigación considerando la recolección, procesamiento y análisis de datos.

4. En el capítulo 5, explicamos las características de los datos y los experimentos que se aplicaron sobre ellos.

5. En el capítulo 6, exponemos los resultados de los experimentos, y a su vez, generamos una discusión de ellos.

6. Finalmente, en el capítulo 7, resumimos las conclusiones del trabajo y discutimos futuros líneas de investigación.

Capítulo 2

Marco teórico

En este capítulo, buscamos introducir ciertas definiciones y conceptos de diversas áreas que abarcamos en esta investigación. Se explicarán conceptos básicos de sismos, desde qué es un sismo a cómo se miden. Luego definiremos qué es una red social, en qué consisten, y profundizaremos en una en particular, Twitter, sus características y en qué se diferencia de las demás. A continuación, introduciremos distintos métodos de clasificación. Por último, revisaremos formas para evaluar un método y cómo comparar su eficacia con otros.

2.1. Sismos

Sismo según el CSN se define como "*el proceso de generación de ondas y su posterior propagación por el interior de la Tierra. Al llegar a la superficie de la Tierra, estas ondas se dejan sentir tanto por la población como por estructuras, y dependiendo de la amplitud del movimiento (desplazamiento, velocidad y aceleración del suelo) y de su duración, el sismo producirá mayor o menor intensidad*". En la actualidad existen dos escalas establecidas para cuantificar un sismo, Richter y Mercalli.

Richter es una escala que determina la magnitud de un sismo. Esta medida se define como el logaritmo de la amplitud de la onda generada por el sismo. Se interpreta como la cuantificación de la cantidad de energía liberada. Su rango es de 1 a 10 y cabe destacar, que al ser escala logarítmica la diferencia de energía liberada de dos sismos que se diferencian en un punto Richter, es exponencial. La Tabla 2.1, muestra cómo escalan los daños asociados a un sismo según su magnitud y su ocurrencia al año. Un factor que influye en los efectos generados por un sismo, es la ubicación del hipocentro de la superficie. El hipocentro también conocido como foco sísmico, es el punto dentro de la Tierra donde comienza el movimiento sísmico. Además existe el epicentro que corresponde al punto en la superficie situado directamente encima del hipocentro.

La intensidad de **Mercalli** es el efecto que causa un sismo en la superficie, la cual, es un ranking de doce niveles, basado en efectos observados. Para un mismo temblor, normalmente se reportan varias intensidades, las que en general decrecen a medida que la distancia al

Magnitud Richter	Probables efectos	Ocurrencia por año
Menos de 2.0	No percibido.	3000000
2.0 - 2.9	Generalmente no percibida, pero registrada.	365000
3.0 - 3.9	A menudo se percibe, pero rara vez causa daño.	49000
4.0 - 4.9	Percibible movimiento en objetos interiores.	6200
5.0 - 5.9	Daños importantes en edificios mal construidos.	800
6.0 - 6.9	Destructivo en áreas de hasta 160 kms.	120
7.0 - 7.9	Graves daños en áreas mas grande que 160 kms.	18
8.0 - 8.9	Graves daños en cientos de kms.	1
9.0 - 9.9	Devastación en varios miles de kms.	0.05
10+	Nunca registrado.	Extremadamente raro

Tabla 2.1: Escala Richter

epicentro aumenta.

Grado Mercalli	Probables efectos
I	No se advierte a excepción de pocas personas.
II	Se percibe por algunas personas, principalmente que se encuentran en pisos altos.
III	Se percibe en casas y edificios.
IV	Oscilan los objetos colgantes.
V	La mayoría de las personas lo percibe.
VI	Todas las personas lo perciben y sienten inseguridad al caminar. Los objetos se desplazan o vuelcan.
VII	Se experimentan dificultades para mantenerse en pie. Se pueden producir daños en estructuras mal construidas.
VIII	Se realizan daños considerables a estructuras y posibles derrumbes parciales de estos.
IX	Pánico generalizado. La mayoría de los edificios sufre grandes daños. La tierra se fisura.
X	Se destruye gran parte de las estructuras.
XI	Muy pocas estructuras quedan en pie.
XII	Daño prácticamente total. Desplazamiento de enormes rocas.

Tabla 2.2: Escala Mercalli

El Mercalli más grande que se ha registro, fue de una intensidad entre XI y XII en las cercanías de Lumaco, provincia de Malleco, Región de la Araucania, al ser el epicentro del terremoto de Valdivia de 1960.

2.2. Redes sociales

Los servicios de red social o medios sociales son plataformas donde el contenido es generado por los mismo usuarios. Un usuario puede ser un individuo y/o organización. En estas plataformas se crea un perfil por usuario y se permite la interacción entre usuarios.

Twitter es una red social que se caracteriza por los breves mensajes para comunicarse, con un máximo de 140 caracteres[1], denominados *tweets*. En esta red social, existe una relación asimétrica entre los usuarios, al ser optativas y se ven reflejados en el concepto de *seguidores* y *quienes sigues*. No es necesario el consentimiento mutuo entre los usuarios para conocer la información, basta ser el *seguidor* de un usuario para ver su información publicada, independiente que el otro usuario lo siga o no. Además, presenta otras características destacadas como el uso de *hashtag* (símbolo #) antes de una palabra clave, lo que permite el seguimiento de temas relevantes en la red social o el uso de *mención* (símbolo @) antes de un nombre de usuario, con el fin de notificar a dicho usuario del tweet.

Cada usuario de Twitter tiene un perfil, que presenta información relevante del usuario. Dentro de esta información destaca, la cantidad de seguidores, cantidad de personas que sigue, ubicación geográfica, cantidad de tweets que ha realizado, descripción del usuario, foto de perfil, entre otros.

Twitter, con el objetivo de compartir su información, dispone de una API[2] que entrega acceso a los datos públicos generados en Twitter. Estos datos son los proporcionados por los usuarios de Twitter, que decidieron hacer públicos. Para la comunicación están definidos *endpoints*[3], divididos en cinco grupos:

- **Cuentas y usuarios**: Permite la administración, mediante la programación, del perfil y configuración de una cuenta. Dentro de las acciones, se encuentran el silenciar o bloquear un usuario, administrar usuarios y seguidores, entre otros.

- **Tweets y respuestas**: Permite los tweets y respuestas como datos públicos. Permitiendo a los desarrolladores el crear Tweets mediante la API o consultar una muestra de los tweets a partir de palabras claves y/o usuarios específicos. Estos datos, permiten comprender distintos aspectos de las personas, como los temas de interés, opiniones y comprender la información que se difunde por los usuarios.

- **Mensajes directos**: Provee acceso a las conversaciones (*Direct Message*) de usuarios con explícitos permisos concedidos a específicas aplicaciones. Un ejemplo de uso es la creación de conversaciones de empresas con sus clientes, mediante chatbot, logrando mejorar experiencias de servicio al cliente y marketing.

- **Ads**: Permite la creación y administración de campañas publicitarias de forma automática. Mediante los Tweets públicos se pueden identificar temas e intereses, información

[1] El 2017, Twitter cambio la restricción de caracteres por tweet de 140 a 280.

[2] API: métodos definidos para la comunicación entre componentes. La sigla corresponde a *Application Programming Interface*.

[3] Dirección para acceder a un tipo específico de información proporcionada.

importante para definir audiencias específicas de campañas publicitarias.

- **Herramientas de editor y SDKs**: Disponen herramientas para insertar líneas de tiempo, compartir botones y otros contenidos de Twitter en páginas web. Esto permite a las empresas mejorar la experiencia web y facilitar que sus clientes compartan información de sus sitios.

Importante considerar que Twitter restringe las consultas a su API por intervalos de 15 minutos. Dependiendo de la consulta puede variar entre 15 a 180 consultas por intervalo de tiempo.

2.3. Teoría de Grafos

Los grafos son estructuras definidas por un par (V, E), donde V es un conjunto no vacío de vértices y E es un conjunto de arcos, donde un arco es un par de vértices, (x, y) en E. Hay dos tipos de grafos, los dirigidos que significa que las relaciones de los arcos son unidireccional y los no dirigidos que son bidireccionales. Los grafos presentan distintas propiedades, para esta investigación abarcamos las siguientes:

- Un grafo G es denominado **conexo**, si existe un camino entre cualquier par de vértices distintos en G.
- **Componente conexa** de un grafo G es un subgrafo de G que es conexo maximal. Notar que todo grafo G que no es conexo, es la unión de un conjunto de componentes conexas de G.
- La **distancia** entre dos vértices de un grafo, es el número menor de arcos de un camino entre ellos.

2.4. Clasificadores

Un clasificador, como lo expone P. Domingos [10], es un sistema que presenta datos de entrada y salida. Los datos iniciales denominados características, son usualmente representados como un vector discreto o continuo $(x = x_1, x_2, ..., x_{n-1}, x_n)$, a diferencia del valor de salida que es discreto, (y) denominado como clase. El objetivo del clasificador es predecir el valor de y, a partir, del conjunto de características, es decir, identificar el conjunto de características con una clase.

Hay muchos algoritmos de clasificación, pero independiente del algoritmo, todos buscan una función $h(x)$ conocido como función de hipótesis, que busca transformar el vector de entrada x al valor correcto de y. Esta función busca asemejarse lo más similar a la función objetivo del problema de clasificación.

Los clasificadores se puede diferenciar en supervisados y no supervisados. Los supervisados

son aquellos que utilizan un conjunto de entrenamiento, previamente etiquetados por su clase, y un conjunto de evaluación, que mide que tan bueno es el modelo. Por otro lado, existen los clasificadores no supervisados, los cuales no disponen de un conjunto de entrenamiento etiquetado y se basan en las características del conjunto para realizar agrupaciones, que serían las posibles clases.

En esta investigación utilizamos clasificadores supervisados, debido a la existencia de conjuntos de entrenamientos. A continuación explicamos los clasificadores empleados. Para más información sobre los clasificadores revisar [11].

2.4.1. Naïve Bayes

Es un clasificador probabilístico supervisado, que se basa en el Teorema de Bayes para estimar $P(y_i|x)$ más probable, que describa el conjunto $x = (x_1, x_2, ..., x_n)$, donde $y_i \in Y$, siendo Y el conjunto de clases posibles del problema que se pretende clasificar. Lo que se expresa en la ecuación

$$y = argmax_{y_i \in Y} P(y_i|x) \tag{2.1}$$

Aplicando el Teorema de Bayes sobre el lado derecho de la ecuación 2.1.

$$y = argmax_{y_i \in Y} \frac{P(x|y_i)P(y_i)}{P(x)} \tag{2.2}$$

se puede observar que $P(x)$ es un valor constante para cualquier valor de y, ya que x es un dato conocido, por lo que, bastaría analizar el numerador.

$$y = argmax_{y_i \in Y} \frac{P(x|y_i)P(y_i)}{P(x)} \propto argmax_{y_i \in Y} P(x|y_i)P(y_i) \tag{2.3}$$

Notar que $P(x|y_i)$ depende de la distribución conjunta de x dado y_i, por lo que aplicamos la regla de la cadena y considerando que la distribución conjunta de $P(x|y)$, se puede expresar como:

$$P(x_1, ..., x_n|y_i) = P(x_1|y_i)P(x_2|y_i)...P(x_n|y_i) = \prod_{j=1}^{n} P(x_j|y_i) \tag{2.4}$$

Notar que se aplica un fuerte supuesto sobre la independencia condicional de la variable x. Para cada x_k es independiente de cualquier otra x_j para $k \neq j$, lo que implica $P(x_k|x_j, y) = P(x_k|y)$. Finalmente el cálculo de $P(y_i|x)$ más probable queda de la siguiente forma:

$$y = argmax_{y_i \in Y} \frac{P(x|y_i)P(y_i)}{P(x)} = P(y_i) \prod_{j=1}^{n} P(x_j|y_i) \tag{2.5}$$

2.4.2. Clasificador Support Vector Machine

El Support Vector Machine (SVM) es un modelo de aprendizaje supervisado, que plantea el problema como un conjunto de puntos, que representan el conjunto de entrenamiento, en un espacio dimensional y busca construir un hiperplano (o conjunto de hiperplanos), que separe en forma óptima los puntos que pertenezcan a una clase u otra. A partir del hiperplano creado, se busca predecir a que clase pertenece un nuevo punto. Cabe mencionar que el concepto de óptimo, es mejor definido como separación óptima (margen), lo que se refiere a que el hiperplano tenga el máximo distancia con los puntos que estén más cerca de él.

El hiperplano óptimo se define como $f(x) = \beta_0 + \beta^T x$. Recordando que el objetivo es maximizar el margen entre las dos clases que están clasificados los N conjuntos de entrenamiento, este problema es formalmente un problema de optimización:

$$min \frac{1}{2}||\beta||^2 \text{ sujeto a } y_i(\beta^T x_i + \beta_0) \geq 1 \forall i \tag{2.6}$$

2.4.3. Multilayer Perceptron

El Multilayer Perceptron (MLP) es una red neuronal que contiene un conjunto de nodos como fuente (características) denominado *capa de entrada*, uno o mas capas ocultas de nodos (perceptrones) y una capa de salida que representaría la clase resultante. El perceptron computa un valor de salida desde múltiples valores de entrada en una combinación lineal con los *pesos*, luego el resultado se aplica una función de activación no lineal. Un perceptron se expresa matemáticamente como

$$y = \varphi(\sum_{i=1}^{n} w_i x_i + b) \tag{2.7}$$

donde w es el vector de pesos, x el vector de valores de entrada, b el bias y φ la función de activación. Para MLP, la función de activación que se utiliza es la función sigmoidea $1/(1+e^{-x})$ o la tangente hiperbólica $tanh(x)$. Se eligen estas funciones, debido a que presentan la propiedad de transición suave, en otras palabras, un cambio en el valor de entrada produce un pequeño cambio en la salida.

El valor de *bias* permite mover la función de activación a la derecha o izquierda. Además es independiente el *bias* de cada capa con la anterior. En la Figura 2.2, se observan los 3 tipos de capas que tiene el MLP, enfocándose en las capas ocultas. Cada capa está compuesta por un conjunto de perceptrones que trabajan en paralelo, y el resultado de cada perceptron

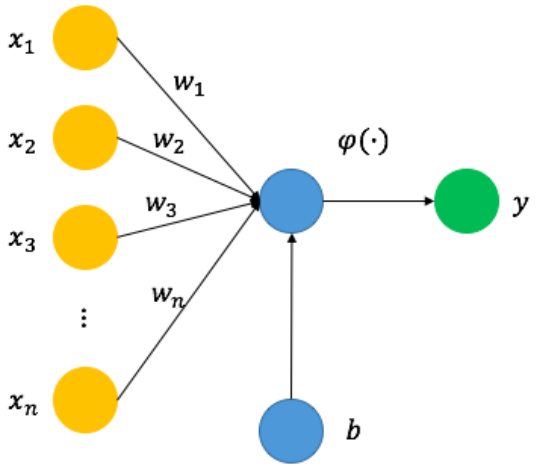


Figura 2.1: Flujo de un perceptron

pasa a ser la entrada de la siguiente capa. La capa de salida realiza una combinación lineal de todas las salidas de la capa anterior.

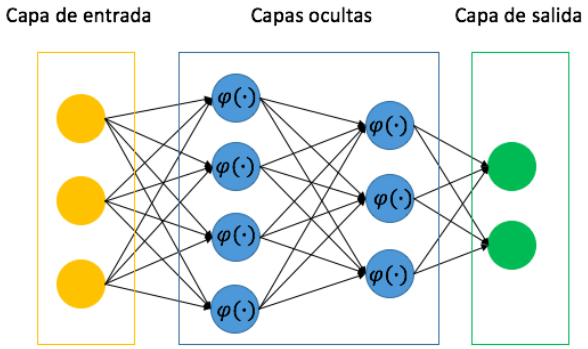


Figura 2.2: Flujo de un MLP

Para aprendizaje supervisado de MLP, se utiliza el algoritmo *back-propagation*. Este algoritmo se divide en dos partes. El ciclo de propagación, donde se evalúa la primera capa con una entrada de prueba propagándose por las siguientes capas hasta generar los valores de salida. A continuación, la etapa de adaptación donde se compara el valor saliente con el ideal, y el error se propaga hacia atrás por las capas. A partir de esta propagación hacia atrás, se va modificando los pesos para disminuir el error. El proceso se repite hasta que los pesos converjan.

2.5. Criterio de evaluación

Al momento de utilizar un clasificador, es muy importante determinar que tan bueno es frente al problema planteado. Aún más relevante es el comparar dos clasificadores para un

mismo problema. Para esto veremos dos criterios de evaluación, contando el costo y evaluación numérica. Se reducirá el problema a un clasificador de dos clases 1 y -1 o positivo y negativo.

2.5.1. Contando el costo

Para un clasificador binario, hay cuatro posibles casos, que se pueden vislumbrar en la Tabla 2.3.

	$y = 1$	$y = -1$
$h(x) = 1$	TP	FP
$h(x) = -1$	FN	TN

Tabla 2.3: Matriz de confusión

Considerando TP (True Positive) las predicciones positivas correctamente clasificadas, FN (False Negative) predicciones negativas incorrectamente clasificadas, FP (False Positive) predicciones positivas incorrectamente clasificadas y TN (True Negative) las predicciones negativas correctamente clasificadas.

A partir de las variables descritas anteriormente se expresan las siguientes métricas de evaluación:

- **Precision**: Esta métrica es la cantidad de aciertos positivos frente al total de predicciones como positivos, lo que se traduce en ¿Cuántos elementos predicho como positivos, son realmente positivos?

$$Precision = \frac{TP}{TP + FP} \tag{2.8}$$

- **Recall**: Esta métrica es la cantidad de aciertos positivos frente al total de elementos que son positivos o ratio de verdaderos positivos (TPR), es decir, ¿Cuántos elementos positivos se lograron predecir del total de positivos?.

$$Recall = \frac{TP}{TP + FN} \tag{2.9}$$

- **FP Rate**: Corresponde a la fracción de incorrectos positivos sobre el conjunto de negativas reales.

$$FPRate = \frac{FP}{FP + TN} \tag{2.10}$$

- **F-measure**: Es el promedio armónico entre Precision y Recall:

$$F - measure = (1 + \beta^2)\frac{2 * Precision * Recall}{\beta^2 * Precision + Recall} \tag{2.11}$$

- **Curva ROC**: Es el gráfico del ratio de verdaderos positivos (TPR) contra el ratio de falsos positivos (FPR). A partir de este gráfico se pueden extraer varias métricas estadísticas, pero la que es utilizada en el aprendizaje de máquina es el área bajo la curva ROC, que se interpreta como la probabilidad que un clasificador elija aleatoriamente más alto, una instancia positiva que una negativa.

2.5.2. Evaluación

Anteriormente se revisó el error presente en los resultados como correcto o incorrecto valor predicho, aplicables generalmente a clasificación binaria. En el caso de clasificaciones no binarias, podemos aplicar la evaluación numérica, donde se evalúa que tan lejos estuvo el valor de pronóstico al valor observado. Ahora consideremos X como los valores reales y $\hat{X}$ como los valores estimados. A continuación se observan posibles métricas a utilizar:

- **Mean Absolute Error (MAE)**:

$$MAE = \frac{1}{n}\sum_{\mathrm{i}=1}^{n}|X_{\mathrm{i}} - \hat{X}_{\mathrm{i}}| \tag{2.12}$$

- **Root Mean Squared Deviation (RMSD)**:

$$RMSD = \sqrt{\frac{1}{n}\sum_{\mathrm{i}=1}^{n}(X_{\mathrm{i}} - \hat{X}_{\mathrm{i}})^2} \tag{2.13}$$

Capítulo 3

Trabajo relacionado

En este capítulo, expondremos las investigaciones relacionadas con Twitter y eventos sísmicos. Primero, introducimos las redes sociales como fuente de información en eventos sísmicos. A continuación explicamos que características son extraíbles de las redes sociales para el análisis sobre un evento físico. Finalmente, discutimos trabajos previos que apliquen análisis espacio-tiempo con el objetivo de hacer predicciones usando redes sociales.

3.1. Uso de Twitter como sensor social

Twitter es una red social 2006, que permite a los usuarios escribir mensajes denominados tweets, con un máximo de 140 caracteres[1]. Presenta 300 millones de usuarios activos al mes aproximadamente, los cuales generan información diariamente en esta red social. Por lo anterior, esta red ha sido tema de investigación en el ámbito de la información a lo largo de estos años. Se han realizados trabajos relacionados con la búsquedas de patrones y predicción temprana de eventos a partir de la red social.

Ahora bien, con toda la información que se crea diariamente, surge la incertidumbre de lo fidedigna que puede ser esta, ya que se puede crear contenidos falsos de la realidad y al no existir moderador, no habría validación de los datos. Gracias al trabajo de C. Castillo et al. [1], relacionado con la credibilidad de la información en Twitter, se sabe que en situaciones de emergencia como un desastre natural, existe una diferenciación entre la información real y falsa. Esto se evidenció en la forma de propagación de la información entre los usuarios, donde se encontraron patrones que distinguen lo real de lo falso. A partir de esto, es posible aventurarse en las redes sociales, como una fuente confiable a utilizar para conocer y entender un evento físico.

[1]En Noviembre del 2017 se modificó la limitación de caracteres a 280. Para esta investigación no aplica dicha modificación

3.1.1. Sensor social

Es importante entender el concepto de *sensor social*, el cual consiste en usuarios que generan información en redes sociales [12]. Esta información, se compone de los gustos del usuario, su ubicación, sus pensamientos, la descripción de su entorno, entre otros. En paralelo, T. Sakaki et al. [3] utiliza el término de sensor social al describir su fuente de datos en la detección de sismos. Considera a los usuarios de Twitter como sensores y los tweets como datos extraídos del sensor. Además introduce el concepto de sensor inoperable, el cual consiste en usuarios que están ocupados realizando otras actividades, lo que puede ser considerado ruido en los datos.

3.1.2. Twitter en desastres naturales

A partir de lo anterior, surge la idea de que las redes sociales, pueden informar de lo acontecido en la realidad, en casos extremos, como desastres naturales. Un estudio de M. Mendoza et al. [13] para el terremoto del 2010 en Chile, demostró que Twitter puede establecer la información que es real o cuestionable, en otras palabras, podemos diferenciar la información confiable durante una emergencia. En el trabajo mencionado, se utilizó la forma de propagación de la información para encontrar este comportamiento. Incluso en Europa, se ha demostrado que frente a desastres naturales, con la información de los usuarios de Twitter, se ha logrado mapear las zonas de impacto de la emergencia [14]. Entonces, nos da la posibilidad de usar las redes sociales para comprender la situación en eventos físicos.

Detección de eventos físicos

A. Hughes et al. [15] plantea que frente a un evento masivo o emergencia es posible identificar un comportamiento característico de los usuarios de Twitter que permitiría diferenciar estos eventos con acontecimientos diarios. J. Bagrow [16] muestra la forma en que las personas propagan la información en emergencias de gran escala (apagones, alerta de bombas, terremotos, entre otros). En adición, los datos provenientes de las redes sociales han servido en distintos eventos físicos como el análisis de incendios forestales [17], accidentes de buses en el transporte público [18] y caídas en centrales eléctricas [19]. Se ha profundizado como línea de investigación en redes sociales, la alerta temprana de un evento físico, a través de los sensores sociales, como el logrado por R. Li et al. [20], donde crearon un sistema que permitió identificar de forma temprana, crímenes y desastres de una localidad. También T. Sakaki et al. [3], realizó un sistema para la detección de sismos logrando detectar la localización del evento a partir de los datos entregados por Twitter.

J. Guzmán y B. Poblete [21] usan el concepto de *burst*, definido como un gran número de eventos ocurridos en una ventana de tiempo [22]. En particular, los autores definen el concepto *bursty keyword*, que consiste en una palabra o conjunto de palabras que repentinamente se presentan en una secuencia de texto a una tasa de aparición inusualmente alta [23]. Los autores proponen un detector de sismos basado en *Bursty Keyword*. El detector se caracteriza por la escalabilidad y eficiencia del algoritmo, basado en ventanas de tiempo, y por el uso de

técnicas de eliminación de palabras irrelevantes en una secuencia de texto.

Twitter en sismos

Dentro de los desastres naturales, los sismos son uno de los eventos de emergencia que ha generado mayor interés en estudiar su vínculo con Twitter. TwiFelt [24], es un sistema en Italia que logra estimar la extensión del área percibida por las personas en un terremoto, sin embargo presenta una limitación relacionada con la dependencia de tweets geolocalizados para la eficacia del sistema, es decir, usuarios que utilicen el GPS en Twitter. En Chile el uso de GPS alcanza solo un 8 % de los usuarios. En paralelo, la plataforma ESA [25], sistema que utiliza los mensajes de Twitter para informar la situación de los incidentes de emergencia durante el desarrollo de esta, ha demostrado su eficacia para la detección de sismos [9].

3.2. Caraterísticas extraibles de Twitter

Se han planteado importantes desafíos en la extracción de características, a partir de los datos que se generan en Twitter. Destacamos, el identificar y conectar dichas características con eventos del mundo real.

Existen muchas características posibles de extraer en Twitter como observamos en la Tabla 3.1. Estas características pueden ser categorizadas según su alcance. En el trabajo actual, utilizamos características proveniente de dos de estas categorías, mensaje y propagación.

3.2.1. Características del mensaje

Estas características provienen del contenido del mensaje, en el caso de Twitter, la información de los tweets. M. Naaman et al [26], logran categorizar a ciertos usuarios a partir del contenido que generan en la red social. Esto muestra la posibilidad del contenido del mensaje de un tweet, para inferir nueva información, caracterizar al usuario, clasificar el contenido, entre otros.

En esta categorización de características, existen algunas que son independientes de Twitter, como la cantidad de caracteres o palabras empleadas en un mensaje o el uso o no uso de signos de exclamación. Por otro lado, se han estudiado características que son dependientes de Twitter o exclusivas de este, como son el uso de hashtag o si corresponde a un re-tweet. En este último caso se han realizado investigaciones para predecir cuando un tweet será retuiteado [27] [28].

Otro fenómeno que ha surgido en las redes sociales es el uso de emoticons, que consisten en tipografías creadas al combinar caracteres alfanuméricos, como por ejemplo :-), el cual

[2]Cantidad de grupos en una red que no tienen conexión entre ellos. En teoría de grafos, es el número de componentes conexas de un grafo.

Categoría	Característica	Descripción
Léxica	Número de caracteres	Largo del texto de un tweet, en cantidad de caracteres.
Léxica	Número de URLs	Número de URLs que contiene un tweet.
Léxica	¿Es Retweet?	Contiene "RT" en el tweet.
Usuario	Cantidad de seguidores	Número de personas que siguen al autor, al momento de la publicación.
Usuario	¿Esta verificado?	1, si el autor tiene su cuenta verificada.
Usuario	¿Tiene URL?	1, si el autor tiene definido el campo URL de la página de inicio.
Tema	Fracción de retweets	Fracción de los tweets que son retweet.
Tema	Largo promedio	Promedio del largo de un tweet.
Tema	Cantidad de distintos hashtags	Número de distintos hashtags.
Propagación	Número de semillas	Cantidad de semillas.[2]
Propagación	Número de semillas no aisladas	Cantidad de semillas (sin semillas de un único usuario).
Propagación	Promedio de usuarios en semillas	Promedio de usuarios en semillas.

Tabla 3.1: Características extraídas de Twitter, expuestas por C. Castillo et al. [1]. Divididas en 4 categorías, según el origen de las características.

corresponde a una cara sonriendo que representa felicidad. Estos elementos en un texto pueden describir sensaciones que esta sintiendo el usuario de la red social, y puede ayudar a caracterizar lo que está sucediendo en un evento físico, como es expuesto en [29].

3.2.2. Características de la propagación

Un aspecto interesante de las redes sociales viene en parte por su definición, "*un conjunto bien delimitado de actores, individuos, grupos, organizaciones, comunidades, sociedades globales, - vinculados unos a otros a través de una relación o un conjunto de relaciones sociales*" (Lozares 1996:108). Este definición permite que se pueda extraer información que no provenga del texto de un tweet, sino de las relaciones entre usuarios. En el caso de Twitter, los usuarios se conectan entre sí, mediante el seguimiento de un usuario a otro, llamados seguidor y seguido respectivamente. Esta cualidad permite al seguidor leer los tweets que escriben los usuarios que está siguiendo, por consiguiente, un usuario al tener más seguidores, se considera un ente más relevante o influyente en la red, ya que sus tweets son leídos por más personas que otro usuario que tenga pocos seguidores. Estas interacciones entre usuarios de Twitter se pueden representar con un grafo dirigido, debido a que si el usuario A sigue al usuario B, no necesariamente el usuario B sigue a A.

Se extraen las siguientes características de la red de Twitter:

- **Cantidad de semillas**: Es la cantidad de grupos que no tienen conexión entre ellos, lo que se traduce en un grafo al número de componentes conexas.

- **Cantidad de usuarios en semillas**: Es la cantidad promedio de usuarios que pertenecen a las semillas, lo que se traduce en un grafo al promedio de vértices en las componentes conexas.

- **Profundidad de la semilla**: Consiste en encontrar el número promedio o máximo de distancia entre los usuarios por semilla, lo que se traduce en la distancia promedia o máxima de los vértices de cada componentes.

3.3. Análisis espacio-temporal

El gran potencial de Twitter en el área científica como objetivo de investigación. Un aspecto a profundizar es la geo-referenciación, un dato que algunas redes sociales lo manejan, en particular, Twitter. E. Steiger et al. [30] revisaron las actuales investigaciones de métodos y aplicaciones del análisis espacio-temporal en Twitter, mostrando una prometedora oportunidad en la GIScience[3], pero a su vez, un campo aún inexplorado, que permitiría entender los procesos geográficos y relaciones espaciales dentro de la red social.

Un primer elemento a considerar, es la extracción de la geo-localización de un usuario. En pocos casos la información es entregada por el usuario como dato de su perfil de la red social o información de coordenadas GPS del mismo tweet. Destacamos la existencia de una importante correlación entre ambos atributos espaciales [31], pero con una baja cobertura del universo de usuarios. Entonces surge la problemática, ¿Podemos inferir la ubicación geográfica?. TweoLocator [32], logró en cierta medida inferir la localización, la cual, se basa exclusivamente en el contenido del tweet para identificar su ubicación. Por otro lado, C. Davis et al. [33] exponen un método para inferir la localidad a partir de las relaciones existentes entre los usuarios.

Además se han realizado trabajos sobre la visualización de estos atributos, por ejemplo D. Thom et al. [34], lograron mostrar un análisis de la información frente a un evento anómalo, incluso diferenciando si corresponde a un evento local o global.

Profundizando aún más en el uso de lo espacio-temporal, algunos autores lo incorporaron como parte de las características de modelos predictivos, como es el caso de detección de sismos [3] y la caracterización de la estructura en las actividades urbanas [35].

[3]Es el análisis, almacenamiento y visualización y gestión de datos geográficos

Capítulo 4

Metodología

En este capítulo, describimos el procedimiento para utilizar la información de las redes sociales, en particular Twitter, para inferir intensidad de Mercalli. Este proceso se divide en 3 etapas: extracción de las características, estimación de Mercalli y estimación espacial de Mercalli.

La primera etapa comprende desde la extracción de la información proveniente de Twitter hasta la generación de las características que definirán un evento sísmico. Cada evento será caracterizado a nivel de distritos geográficos, en este caso utilizamos comunas de todo Chile. Esta agregación geográfica será muy relevante para el método que se utiliza en la tercera etapa. La segunda etapa, se divide en dos procesos consecutivos, por un lado, predecir las zonas que perciben el sismo y a continuación, aplicar sobre estas zonas un modelo de regresión para inferir la intensidad de Mercalli a nivel de comuna. En la tercera y última etapa, se utilizan los puntos estimados de la etapa anterior para realizar una predicción espacial de Mercalli.

4.1. Extracción de características

Recolectamos toda la información de los tweets que mencionen algunas de las palabras claves asociadas a eventos sísmicos, como "*temblor*", "*sismo*" o "*terremoto*". Consideramos una ventana de tiempo de 30 minutos desde que ocurrió el evento para recolectar los tweets.

Los tweets son agrupados según comuna, por lo que es necesario identificar la ubicación geográfica de estos. Recuperamos la ubicación, mediante varias fuentes de información que se procesan de forma secuencial. En caso que la primera fuente no contenga dicha información, se procede con la siguiente, y así sucesivamente. El orden de procesamiento para identificar la localización se muestra a continuación:

1. Revisar el campo de coordenadas que proviene del tweet.

2. Utilizar el contenido del mismo de tweet. Como contiene lenguaje natural se puede

recuperar la comuna a partir de este texto natural aplicando el método *fuzzy string matching*[1].

3. Utilizar la ubicación del perfil del usuario y de igual forma, aplicar el método *fuzzy string matching*

Se definió un conjunto de 19 métricas que se especifican en la Tabla 4.1. Las métricas identifican la percepción de la gente en un sismo o terremoto, y son calculadas a nivel de comuna para caracterizar el evento sísmico. Estas características están inspiradas en el trabajo de C. Castillo et al. [1]

[1]Método de búsqueda de palabras que coincidan aproximadamente con un patrón definido.

Categoría	Característica	Descripción
Léxica	N de Tweets	Cantidad de tweets en un sismo.
Léxica	Tweets Normalizados	Fracción de tweets sobre la población de la ciudad.
Léxica	Promedio de palabras	Promedio de la cantidad de palabras de los tweets.
Léxica	Promedio de caracteres	Promedio de la cantidad de carácteres de los tweets.
Léxica	Signos de pregunta	Fracción de los tweets que contienen el signo '?'.
Léxica	Signos de exclamación	Fracción de los tweets que contienen el signo '!'.
Léxica	Palabras con mayúscula	Fracción de los tweets con mayúscula.
Léxica	Uso de Hashtag	Fracción de los tweets que contienen el hastag ('#').
Léxica	Uso de símbolo de mención	Fracción de los tweets que contienen el mención ('@').
Léxica	Símbolo de retweet	Fracción de lo tweets que contiene 'RT'.
Léxica	Contiene la palabra 'terremoto'	Fracción de los tweets con la palabra "terremoto".
Red	N de semillas	Cantidad de semillas.
Red	N de semillas no aisladas	Cantidad de semillas (sin semillas de un único usuario).
Red	Promedio de usuarios en semillas	Promedio de cantidad de usuarios existentes en cada semilla.
Red	Promedio de usuarios en semillas (sin solitarias)	Promedio de cantidad de usuarios existentes en cada semilla (semillas no aisladas).
Red	Semilla más grande	La semilla que contiene más usuarios.
Red	Promedio de conexiones en una semilla	Promedio de conexiones de usuarios en la semilla.
Red	Promedio de conexiones en una semilla (sin solitarios)	Promedio de conexiones de usuarios en la semilla (semillas no aisladas).
Red	Máximas conexiones de una semilla	La semilla que tienen más conexiones entre usuarios.

Tabla 4.1: Métricas que se dividen en 2 categorías. Métricas léxicas, que se relacionan con el contenido del tweet. Métricas de red, que van asociados a las relaciones que presentan los usuarios que generan los tweets.

4.2. Estimación de Mercalli

Buscamos identificar y separar las comunas donde el evento es percibido, denominadas regiones de interés, de las comunas que fue comentado el evento pero no fue percibido, denominadas regiones no relevantes. Observamos que la regiones no relevantes no presentan intensidad de Mercalli, al ser comunas que no perciben el evento, por lo que al excluirlos se evita en el modelo de estimación de Mercalli un error proveniente de estas comunas que teóricamente su intensidad de Mercalli corresponde a cero. Las regiones de interés serán utilizadas para predecir la intensidad estimada de Mercalli, usando un clasificador $0/1$, entrenado sobre los sismos de la ventana de tiempo definido anteriormente. Etiquetamos los datos acorde al reporte de intensidad de Mercalli en dos clases disjuntas. La clase 0 representa un sismo que no ha sido percibido por la población de la zona (sin reporte en la escala de Mercalli) y la clase 1 representa un sismo que ha sido percibido con un valor definido de intensidad en la escala de Mercalli. Cada conjunto es separado según evento y comuna, donde es representado por un vector de características que se definieron en la Tabla 4.1. Una vez, el clasificador $0/1$ es entrenado, el método está listo para detectar las comunas de interés en un nuevo conjunto de sismos.

Sobre el conjunto de comunas etiquetadas con la clase 1 aplicamos un modelo para estimar la intensidad Mercalli por cada comuna. Estos resultados serán usados como datos de entrada para caracterizar el sismo en la última etapa del proceso, que corresponde a la estimación espacial de Mercalli.

Indicar que este procedimiento busca separar las comunas donde el evento fuer percibido de aquellas en las cuales fue comentado pero no percibido

4.3. Estimación espacial de Mercalli

Este método busca mejorar la estimación de la intensidad en la escala Mercalli en el modelo de regresión. Esto se lleva a cabo, al incluir la dimensión espacial en el modelo. Predecimos la intensidad de Mercalli de una zona o comuna considerando las estimaciones de las zonas cercanas, en otras palabras, las zonas vecinas de la comuna influirán en la predicción.

Primero tenemos que calcular el Mercalli reforzado a nivel de comuna. Considérese i el índice de una comuna del conjunto de comunas de interés de un sismo. Definimos soporte local $s(\mathrm{i}) \in [0, 1]$ para el i-th comuna, como la proporción entre usuarios de la comuna i que reportaron el sismo y el número total de usuarios de Twitter de la comuna i. Sea $m(\mathrm{i})$ la intensidad estimada con el modelo de regresión. Combinamos $m(\mathrm{i})$ y $s(\mathrm{i})$ para generar el Mercalli reforzado. Los puntos de estimación de intensidad con bajo soporte serán descartados al combinar ambos factores en la siguiente función:

$$\text{Mercalli Reforzado}(\mathrm{i}) = m_{reforzado}(\mathrm{i}) = \frac{2 \cdot \overline{m}(\mathrm{i}) \cdot s(\mathrm{i})}{\overline{m}(\mathrm{i}) + s(\mathrm{i})}, \tag{4.1}$$

donde $\overline{m}$(i) es la intensidad estimada de la comuna i en el intervalo $[0, 1]$. Para esto, se normalizó como $\overline{m}(\mathrm{i}) = \frac{m(\mathrm{i})-1}{11}$. De esta manera $\overline{m}$(i) y s(i) están en el mismo rango $[0, 1]$, y a su vez, la función Mercalli Reforzado está en el rango $[0, 1]$. Mostramos un diagrama de los niveles de contorno de la función en la Figura 4.1.

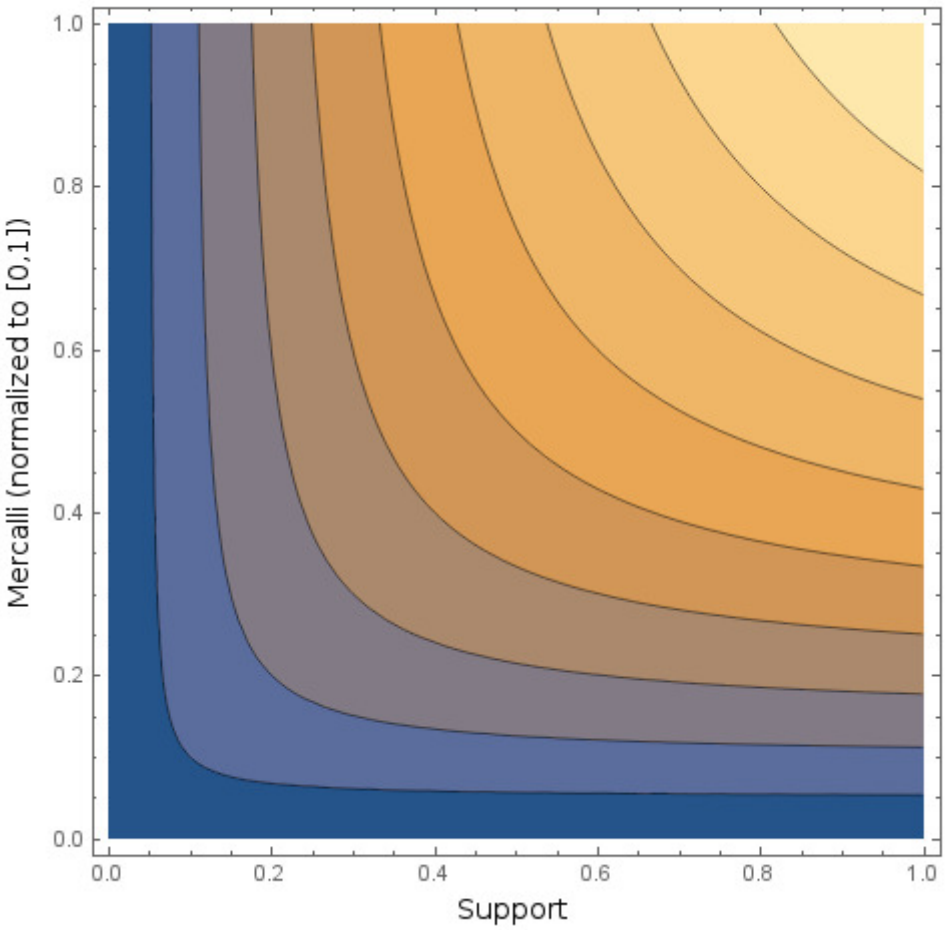


Figura 4.1: Función del Mercalli Reforzado usando niveles de contorno.

Mercalli Reforzado busca darle mayor relevancia a las comunas con más soporte en Twitter sobre el sismo ocurrido, para así disminuir los falsos positivos. Esto se traduce en considerar los datos con alto soporte.

El Mercalli Ajustado considera cada comuna como un sensor. La reacción de las personas frente a un sismo es una señal que se analizará por comuna. Para modelar la activación de nuestro sensor usamos una función de activación. Tomamos una función sigmoide $\frac{1}{1+\mathrm{e}^{-x}}$ para modelar el nivel de activación de una comuna frente a un evento sísmico. La activación de la función para la intensidad de Mercalli sera fijado en 3, que es definido como el primer nivel de la escala de Mercalli donde es sentido un sismo. Entonces, una comuna es considerada como activa desde nivel 3 hacia arriba donde el sismo será reportado. Para aplicar la función sigmoide en el Mercalli Reforzado, volvemos a la escala $[1, 12]$ y desplazándolo a 3, es decir, $(11 \cdot m_{reforzado}(\mathrm{i}) + 1) - 2$. Denotamos $M_{adj}(\mathrm{i})$ para referirnos al Mercalli Ajustado,

siendo la combinación de la función sigmoide que esta en el rango de $[0, 1]$, con el Mercalli estimado de la comuna i, lo que correspondería a la siguiente expresión:

$$M_{adj}(\mathrm{i}) = m(\mathrm{i}) \cdot \textsc{Sigmoide}\left(11 \cdot \left[\frac{2 \cdot \overline{m}(\mathrm{i}) \cdot s(\mathrm{i})}{\overline{m}(\mathrm{i}) + s(\mathrm{i})}\right] - 1\right). \tag{4.2}$$

En la Figura 4.2 visualizamos el comportamiento del Mercalli ajustado que permite enfocar el trabajo en la detección de sismos con una intensidad mínima.

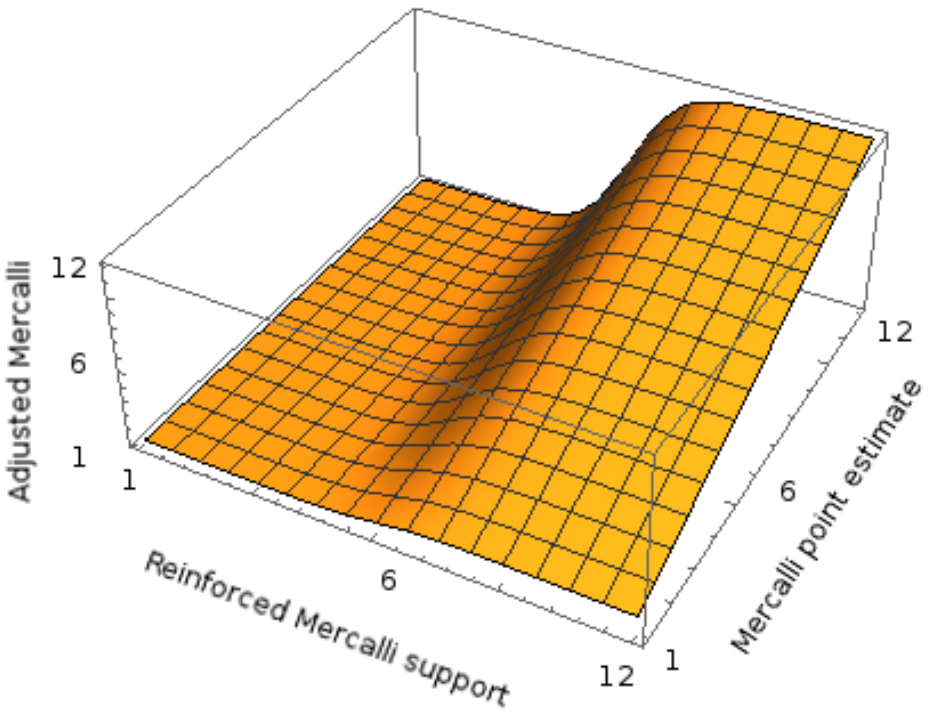


Figura 4.2: La intensidad del Mercalli Ajustado combina los efectos del punto de Mercalli estimado y el soporte en una colección de funciones sigmoides. El nivel de activación es un parámetro de nuestra escala de intensidad. Por ejemplo, esta superficie muestra un nivel de activación fijada en 6 sobre la escala de Mercalli, restringiendo el método para la detección de sismos de alta intensidad.

Las comunas son parte de una región geográfica. Es necesario considerar el efecto de las relaciones espaciales con las comunas vecinas. Para hacer esto, suavizamos el Mercalli Ajustado[2] estimado por cada comuna en relación al Mercalli Ajustado estimado de los

[2]Notar que Mercalli Ajustado estima a nivel de comuna como una zona aislada.

vecinos de este. El efecto de cada vecino es inversamente proporcional a la distancia geodésica de la comuna. Calculamos la distancia geodésica entre todos los pares de comunas del evento. Posterior, por cada comuna se extrae la lista de k vecinos cercanos, conocido como k-NN(i), donde k es un parámetro a definir que representa la cantidad de vecinos para la comuna i. Entonces para cada comuna i, normalizamos la distancia a la comuna i de cada k-vecino cercano por la suma de las distancias de la lista k-NN(i), como se expresa a continuación:

$$\mathrm{d}(\mathrm{i}, j) = \frac{\mathrm{d}_{geo}(\mathrm{i}, j)}{\sum_{j' \in k\text{-NN(i)}} \mathrm{d}_{geo}(\mathrm{i}, j')}. \tag{4.3}$$

Damos mayor relevancia a los vecinos que se encuentren más próximos a la comuna i, por lo que expresamos la distancia de la siguiente manera:

$$\mathrm{SIM}(\mathrm{i}, j) = \frac{1 - \mathrm{d}(\mathrm{i}, j)}{\sum_{j' \in k\text{-NN(i)}} 1 - \mathrm{d}(\mathrm{i}, j')}. \tag{4.4}$$

Suavizamos el punto estimado de Mercalli para la comuna i, usando una combinación lineal del punto estimado y las estimaciones de los vecinos de la comuna i.

$$\mathrm{M(i)} = (1 - \lambda) \cdot M_{\mathrm{adj}}(\mathrm{i}) + \lambda \cdot \sum_{j' \in k\text{-NN(i)}} \mathrm{SIM}(\mathrm{i}, j') \cdot M_{\mathrm{adj}}(j), \tag{4.5}$$

donde λ es un parámetro con valores entre $[0, 1]$, que controla el peso que tiene el punto estimado y su vecindad. Notar que $\lambda = 1$ corresponde a que M(i) depende exclusivamente de los vecinos y $\lambda = 0$ significa que M(i) depende solamente por el punto estimado. Como la escala de Mercalli utiliza números enteros, aproximamos M(i) al entero más cercano.

Capítulo 5

Experimentación

En esta capítulo, se detallan los experimentos que se efectuaron con datos de Twitter asociados a eventos sísmicos ocurridos en Chile. Primero, describimos el origen y las características de este conjunto de datos. A continuación, exponemos experimentos aplicados en Richter y Mercalli y comparamos sus comportamientos. Proseguimos, con la generación del modelo de datos. Finalmente, mostramos la aplicación de la estimación espacial planteada en la sección 4.3 sobre el modelo de datos.

5.1. Datos para la experimentación

Según los objetivos planteados, se desea aplicar la investigación en Chile, para así comprender la relación existente entre redes sociales y eventos sísmico en nuestro país. Por ello se emplearon datos relacionados con eventos sísmicos originados en el territorio nacional en un tiempo definido.

La recolección de datos se logró gracias al trabajo realizado por el grupo de investigación PRISMA[1], el cual buscaba mostrar la factibilidad de un modelo de detección de eventos sísmicos [36]. Para verificar, crearon un sistema donde uno de sus componentes consistía en la recolección de tweets para eventos sísmicos utilizando el método *Efficient Bursty Keyword Detection Model* [21]. Este componente se utilizó como fuente de información para los datos de nuestra investigación, recolectando información entre el año 2016 y el primer semestre del año 2017.

5.1.1. Registros de sismos

En el período mencionado anteriormente, se registran 560 sismos sensibles por el ser humano, con un rango de magnitud Richter entre 2.2 Mw y 7.6 Mw. Los sismos ocurridos en

[1]Grupo de investigación enfocado en la minería de datos y bases de datos multimedia

Chile son registrados por el Centro Sismológico de Chile (CSN). El CSN entrega esta información por medio de su página web[2]. Adicionalmente, este organismo provee por cada sismo sensible por el ser humano reportes de Mercalli según las zonas afectadas, lo que contempla 3.602 reportes en el período comprendido.

La información de los sismos, dispuesta en su página web, se visualiza por día, mostrando el listado de todos los sismos ocurridos, incluyendo los sismos sensibles y no sensibles por el ser humano. Cada página que muestra el listado de un día, se representa de la siguiente manera `http://www.sismologia.cl/events/listados/YYYY/MM/YYYYMMDD.html`, donde YYYY representa el año, MM el mes y DD el día. Un ejemplo del listado de sismos que entrega el CSN, es el observado en la Figura 5.1. Cada fila representa un sismo ocurrido en una zona del territorio de Chile. Para identificar los sismos sensibles, se resaltan con una tipografía diferente. Cada sismo sensible contiene un enlace que detalla las características de este evento y los reportes de Mercalli.

Fecha Local	Fecha UTC	Latitud	Longitud	Profundidad [Km]	Magnitud	Referencia
14/04/2017 19:50:50	14/04/2017 22:50:50	-21.711	-68.409	123.6	2.7 Ml GUC	56 km al S de Ollagüe
14/04/2017 17:35:48	14/04/2017 20:35:48	-27.964	-69.273	118.5	3.4 Ml GUC	111 km al SE de Tierra Amarilla
14/04/2017 15:50:40	14/04/2017 18:50:40	-31.484	-71.191	59.3	2.7 Ml GUC	17 km al N de Illapel
14/04/2017 15:03:24	**14/04/2017 18:03:24**	**-22.489**	**-68.818**	**121.8**	**5.2 Mw GUC**	**12 km al E de Calama**
14/04/2017 13:22:19	14/04/2017 16:22:19	-29.944	-71.983	35.1	3.3 Ml GUC	58 km al NO de Tongoy
14/04/2017 13:11:27	14/04/2017 16:11:27	-31.991	-71.704	32.6	3.3 Ml GUC	20 km al O de Los Vilos
14/04/2017 13:02:20	14/04/2017 16:02:20	-26.118	-70.769	50.7	3.5 Ml GUC	29 km al NO de Chañaral.
14/04/2017 13:01:20	14/04/2017 16:01:20	-24.005	-67.384	225.8	3.9 Ml GUC	69 km al SE de Socaire
14/04/2017 11:36:30	14/04/2017 14:36:30	-21.350	-68.785	118.3	2.7 Ml GUC	57 km al O de Ollagüe
14/04/2017 10:21:31	14/04/2017 13:21:31	-19.798	-69.205	99.9	3.5 Ml GUC	59 km al SE de Camiña
14/04/2017 09:02:01	14/04/2017 12:02:01	-30.559	-69.379	27.0	4.5 Mw GUC	124 km al SE de Paihuano
14/04/2017 06:55:55	14/04/2017 09:55:55	-19.796	-69.209	99.1	4.4 Ml GUC	58 km al SE de Camiña
14/04/2017 06:29:22	14/04/2017 09:29:22	-32.431	-71.764	31.2	3.3 Ml GUC	50 km al O de La Ligua
14/04/2017 06:22:55	14/04/2017 09:22:55	-38.470	-74.590	25.3	2.8 Ml GUC	97 km al O de Tirúa
14/04/2017 05:20:03	14/04/2017 08:20:03	-30.781	-71.687	29.4	2.9 Ml GUC	41 km al O de Punitaqui
14/04/2017 04:58:57	14/04/2017 07:58:57	-30.318	-71.530	33.7	3.9 Ml GUC	8 km al SO de Tongoy
14/04/2017 02:48:37	14/04/2017 05:48:37	-21.108	-68.973	109.4	2.4 Ml GUC	45 km al SO de Mina Collahuasi
14/04/2017 02:19:21	14/04/2017 05:19:21	-18.986	-70.758	37.9	2.6 Ml GUC	64 km al O de Cuya
14/04/2017 01:35:36	14/04/2017 04:35:36	-30.646	-71.424	41.9	3.6 Ml GUC	22 km al O de Ovalle
14/04/2017 01:24:33	14/04/2017 04:24:33	-30.669	-71.443	46.3	3.0 Ml GUC	24 km al O de Ovalle
13/04/2017 22:36:22	**14/04/2017 01:36:22**	**-31.769**	**-71.617**	**33.5**	**4.2 Ml GUC**	**19 km al NO de Los Vilos**
13/04/2017 21:35:14	14/04/2017 00:35:14	-31.959	-70.787	111.9	3.0 Ml GUC	24 km al SO de Mina Los Pelambres

Figura 5.1: Listado de sismos ocurridos en Chile el día 14 de Mayo del 2017 proveniente del enlace `http://www.sismologia.cl/events/listados/2017/04/20170414.html`

Debido al patrón existente, es posible extraer la información que consideramos relevante para la investigación, mediante el método *web scraping*[3]. Para ello usamos la librería de Python *Beautiful Soup* [37], un intérprete de documentos HTML y XML que se especializa en la extracción de datos en una página web. Extrajimos la información de sismos en Chile, exclusivamente los eventos sensibles por el ser humano e incluyendo por cada evento, los reportes de magnitud Mercalli asociados a una comuna.

En las Figuras 5.2 y 5.3 mostramos la frecuencia de los registros sísmicos, respectivamente Richter y Mercalli, extraídos por el método mencionado anteriormente. En la Figura 5.2, se observa una mayor concentración de sismos cercanos a una magnitud 4 en escala Richter. Recordando la Tabla 2.1, observamos que las magnitudes mayores a 4, presentan una frecuencia

[2]http://www.sismologia.cl/
[3]Método para la extracción de datos desde una página web

menor en un periodo determinado. Estas consideraciones son aplicables en forma equivalente en la Figura 5.3.

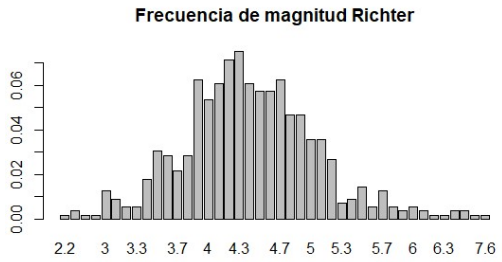


Figura 5.2: Histograma de magnitud Richter en la colección de registros sísmicos

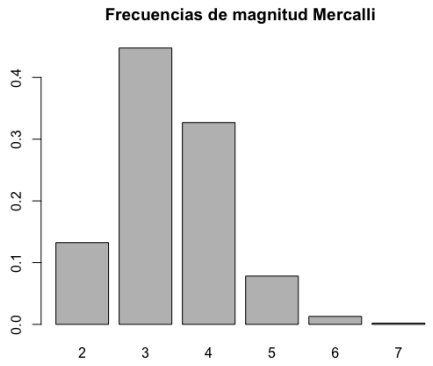


Figura 5.3: Histograma de intensidad Mercalli en la colección de registros sísmicos

5.1.2. Tweets asociados a sismos

En relación a los tweets asociados a sismos, estos cumplen la característica de contener palabras relacionadas a eventos sísmicos, como "*temblor*", "*sismo*" o "*terremoto*" y sus palabras derivados. Los tweets fueron recolectados de una ventana de tiempo de 30 minutos desde el inicio de un sismo. La colección contempla 825.310 tweets, provenientes de 309.750 usuarios únicos distribuidos entre todos los sismos del periodo mencionado.

De la colección de tweets, solamente 2.200 incluyen la información de geolocalización, lo cual representa el 0.26 % de la colección, esta cifra se explica por el poco uso del GPS de los usuarios de Twitter como se mencionó en la sección 3.1.2. Por otro lado, de los 309.750 usuarios únicos, 207.015 registran una localización en su perfil, correspondiendo al 66.8 % de los usuarios. Debido a la baja geolocalización presente en los tweets, inferimos la localización de estos por medio de la ubicación del usuario. De los 207.015 usuarios con ubicación en nuestro conjunto de datos, 57.546 indicaban Chile como país. Con este conjunto se asocio los usuarios con comunas de Chile por medio del método *approximate matching*, implementado en *Fuzzy wuzzy*[4]. Usando un 80 % de nivel de confianza, se obtuvo con este método un total de 41.885 usuarios mapeados con alguna de las comunas de Chile, que corresponde al 72.8 % de los usuarios que indican su ubicación en Chile. Ese conjunto de usuarios, registran 190.249 tweets que ahora están mapeados con 345 comunas distintas de Chile.

En resumen nuestro conjunto de datos Twitter-Sismo, comprende 331 sismos asociados a 190.249 tweets distribuidos en las 345 comunas de Chile por el periodo de 18 meses. Desde la perspectiva de comuna-sismo, se dividen en 6.790 reportes Mercalli percibidos por comuna y 6.548 sismos no percibidos pero con actividad en Twitter relacionado con sismos.

[4]Librería de Python de *string matching* que utiliza la distancia de Levenshtein para calcular la diferencia entre secuencia de letras. Esta disponible en https://github.com/seatgeek/fuzzywuzzy

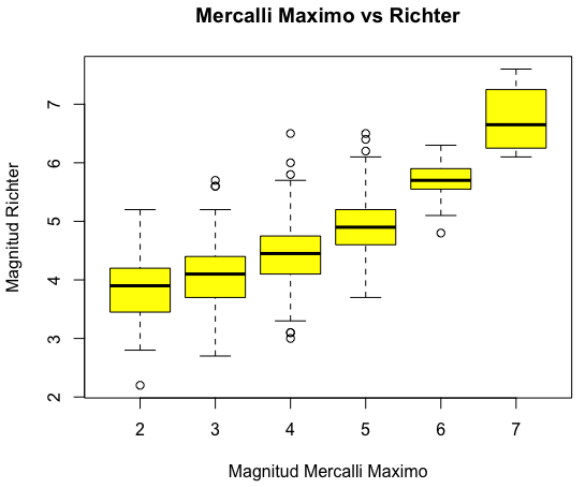
A partir de este conjunto de datos se generan las métricas, mencionadas en la Tabla 4.1, que son divididas en dos categorías, léxicas y de red. La primera categoría necesita un procesamiento del contenido del tweet, aplicando distinta operaciones en el texto. Por otro lado, las métricas de red necesitan la recuperación de información referente a las interacciones que existen entre los usuarios de los tweets del conjunto, en particular, la información de *seguidores* de cada usuario. Para recuperar la información de red, Twitter presenta una API, como se describe en la sección 2.2, que dispone de dicha información. Con esta información de *seguidores*, generamos las métricas de red expuestas anteriormente. Cabe mencionar la dificultad de este proceso, considerando la gran cantidad de usuarios (41.885) y las limitaciones de consultas que permite el API de Twitter en un intervalo definido de tiempo. Debido a lo anterior, fue necesario paralelizar este proceso, expuesto en el apéndice B, y de igual forma provocó un tiempo prolongado en la extracción de datos.

5.2. Mercalli vs Richter

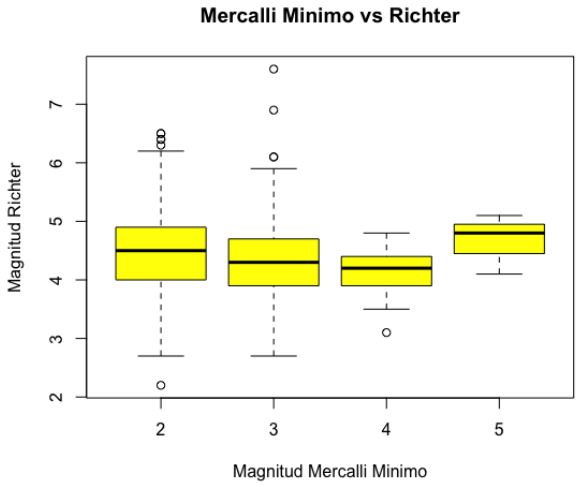
Dentro de esta investigación, determinamos la métrica de un evento sísmico que puede ser caracterizada de mejor manera por las redes sociales. Un sismo es conocido por las personas, a través de dos escalas, Richter y Mercalli. Estos conceptos, expuestos en la sección 2.1, presentan a Richter como una magnitud que caracteriza al sismo por la energía liberada, en cambio, Mercalli o intensidad de Mercalli, es el efecto observable del sismo en una superficie. Una diferencia importante para el estudio es el nivel de desagregación existente entre estas métricas; por un lado, observamos la magnitud Richter como un único valor del sismo, a diferencia de la intensidad de Mercalli, que puede presentar distintos valores en diferentes localidades.

Graficamos ambas métricas, para comprender la relación entre ellas, pero considerando la diferencia existente en el nivel de desagregación en un sismo, definimos generar métricas a partir de la intensidad de Mercalli, que sean a nivel del sismo, las que corresponden a Mercalli máximo, mínimo y promedio de un sismo. Graficamos magnitud Richter según cada una de las nuevas métricas agregadas de Mercalli que observamos en la Figura 5.4.

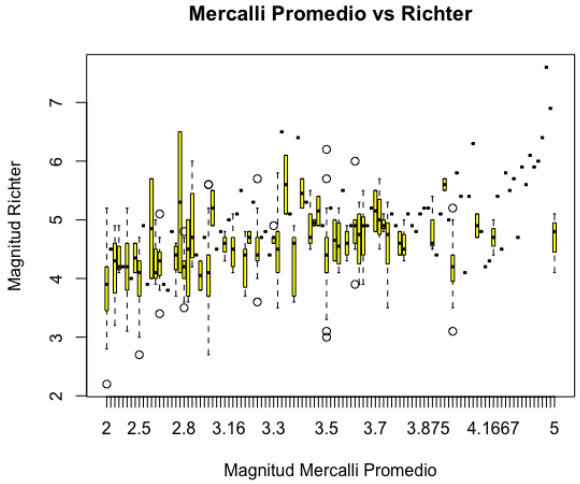
Primero, observamos en la Figura 5.4a que el intervalo que presenta Mercalli máximo abarca todo el espectro de nuestro conjunto de datos y comprobamos la existencia de una relación lineal presente entre Richter con el máximo Mercalli de un sismo. En segundo, la Figura 5.4b a diferencia de la anterior, presenta un rango de Mercalli mínimo, entre grado 2 y 5. Además observamos que el mínimo de Mercalli se concentra entre valores Richter 4 y 5, esto es explicado por la distribución de nuestro conjunto de datos como mostramos en la Figura 5.2. Es importante destacar la ausencia evidente de relación entre Richter y Mercalli mínimo. En tercero y último lugar, en la Figura 5.4c observamos el cambio de discreto a continuo en los valores en el eje Mercalli, entre el rango 2 y 5, debido a que la métrica es el promedio de los valores de Mercalli por sismo. Observamos que no existe una relación entre Richter y el promedio de Mercalli; además es importante considerar esta métrica de Mercalli (promedio de Mercalli), debido a que contempla al conjunto completo de intensidades de un sismo, a diferencia de las anteriores que contempla un único elemento del conjunto.



(a)



(b)



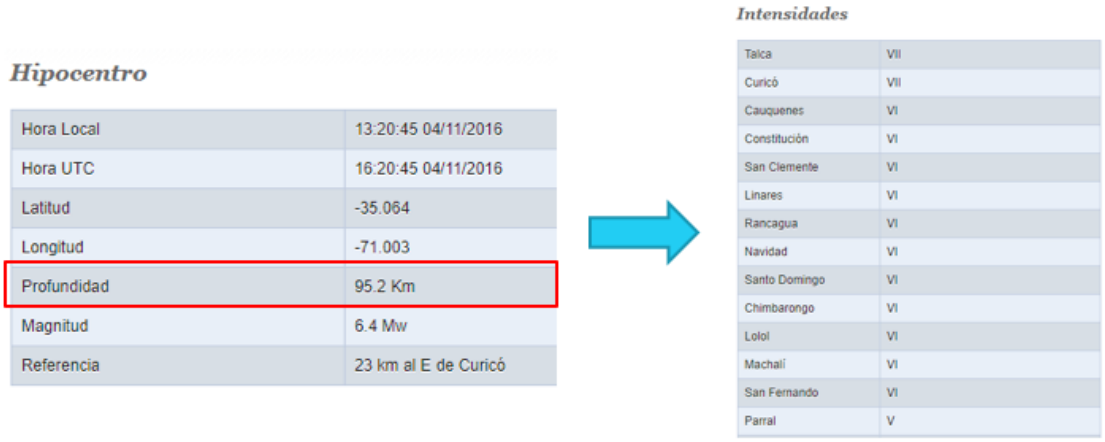
(c)

Figura 5.4: Diagramas de caja por cada tipo de agregación del Mercalli en relación con Richter de un evento sísmico.

Profundizando los comportamientos de magnitudes Richter y Mercalli en un sismo, buscamos un par de sismos con un valor en escala Richter similar, para analizar y comparar los reportes de Mercalli de estos sismos. En la Figura 5.5, presentamos dos sismos con similares magnitudes Richter, 6.4 (Figura 5.5a) y 6.5 (Figura 5.5b) Mw. Podemos observar en la Figura 5.5a, una gran cantidad de reportes de Mercalli y la presencia de altos valores de este mismo. En contra parte, en la Figura 5.5b, presenta una menor cantidad de registros y los valores altos de Mercalli se encuentran muy por debajo en comparación al sismo presentado en la Figura 5.5a.

En ambas figuras remarcamos en rojo una métrica que no habíamos contemplado, la profundidad, que corresponde a la profundidad de la superficie terrestre desde donde se origina un sismo. Además conocemos que los sismos pueden ocurrir en distintas profundidades, independiente de la intensidad en escala Richter. Esto puede ocasionar una variación en las sensaciones de un sismo para el ser humano en la superficie. Podemos reflejar lo anterior, en el caso de la Figura 5.5, donde el sismo de la Figura 5.5a presenta una profundidad significativamente menor a la Figura 5.5b, 95 y 601 metros respectivamente. Evidenciamos que la profundidad es un factor relevante junto a la magnitud Richter del sismo para determinar los reportes de intensidad en escala Mercalli, es decir, manteniendo fijo el valor de Richter en un sismo, mientras menor sea la profundidad del sismo, encontraremos mayores registros de Mercalli y valores más altos.

A partir de lo anterior y para cumplir los objetivos planteados en este trabajo, escogimos para los siguientes experimentos la magnitud de Mercalli. Richter no fue considerado, debido a ser una magnitud única del sismo, que no es posible desagregar a nivel geográfico.



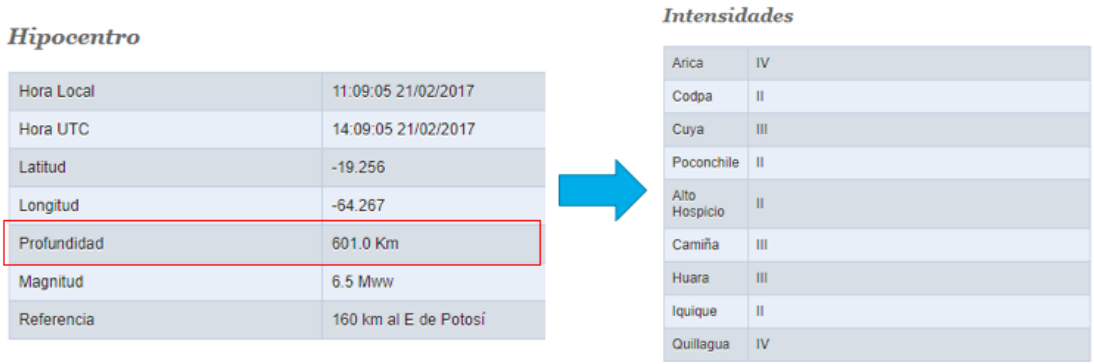
Hipocentro

Hora Local	13:20:45 04/11/2016
Hora UTC	16:20:45 04/11/2016
Latitud	-35.064
Longitud	-71.003
Profundidad	95.2 Km
Magnitud	6.4 Mw
Referencia	23 km al E de Curicó

Intensidades

Talca	VII
Curicó	VII
Cauquenes	VI
Constitución	VI
San Clemente	VI
Linares	VI
Rancagua	VI
Navidad	VI
Santo Domingo	VI
Chimbarongo	VI
Lolol	VI
Machalí	VI
San Fernando	VI
Parral	V

(a) Sismo ocurrido el 4 de Noviembre del 2016



Hipocentro

Hora Local	11:09:05 21/02/2017
Hora UTC	14:09:05 21/02/2017
Latitud	-19.256
Longitud	-64.267
Profundidad	601.0 Km
Magnitud	6.5 Mww
Referencia	160 km al E de Potosí

Intensidades

Arica	IV
Codpa	II
Cuya	III
Poconchile	II
Alto Hospicio	II
Camiña	III
Huara	III
Iquique	II
Quillagua	IV

(b) Sismo ocurrido el 21 de Febrero del 2017

Figura 5.5: Información de Richter y Mercalli de dos sismo ocurrido en Chile. Fuente del CSN

5.3. Generación del modelo

En la generación del modelo, el total de 331 sismos fueron divididos en dos grupos de 263 y 68 eventos, denominado respectivamente conjunto de entrenamiento y prueba. La división de los grupos mantuvo la proporción de 80/20 porcentual del total de sismos. Esta proporción es asignada según el rol que cumple cada conjunto en la generación del modelo, este método denominado muestreo estratificado. El conjunto de entrenamiento corresponde a los eventos sísmicos utilizados para nutrir el modelo. Mientras mayor sea la cantidad de datos de entrenamiento, el modelo caracterizará en mayor profundidad el sismo. Por otro lado, el conjunto de prueba es empleado para validar el modelo, es decir, posterior a la creación del modelo con el conjunto de entrenamiento, aplicamos el modelo sobre el conjunto de prueba. Esta separación, permite verificar que los resultados del modelo no sean ajustados exclusivamente a los datos de entrenamiento, al ser evaluado el modelo posteriormente con el conjunto de prueba. Cabe mencionar que la selección de los eventos entre el conjunto de entrenamiento y prueba, es aleatorio estratificado manteniendo la distribución proporcional de intensidades en escala Mercalli del conjunto total de sismos.

Para la generación y evaluación del modelo utilizamos Weka[5], herramienta para el análisis de datos que contiene la implementación de los distintos clasificadores empleados en este trabajo para la creación de modelos predictivos. Todo modelo de la investigación fue generado usando validación cruzada con 5 iteraciones, técnica de validación de modelo para asegurar la independencia del modelo de los datos.

Como mencionamos en la sección 4.2, creamos en primera instancia un modelo de clasificación 1/0, a partir de los datos de entrenamiento, para identificar las zonas afectadas y no afectadas por un sismo. El conjunto de entrenamiento se compone de 5.021 instancias de la clase 0 (zonas no afectadas) y 5.470 instancias de la clase 1 (zonas afectadas). Probamos para la creación del modelo los clasificadores Support Vector Machine (SVM), Naive Bayes y Multiperceptron, expuestos en la sección 2.4.

Posterior al modelo anterior, continuamos con la creación de otro modelo para predecir el Mercalli a nivel de localidad afectada por el sismo. A partir del conjunto de entrenamiento, seleccionamos exclusivamente las comunas que reportaron intensidad de Mercalli en cada evento. Utilizamos el modelo de regresión de vector de soporte con una optimización mínima secuencial (SMO) y validación cruzada con 5 iteraciones. En el conjunto de datos, existe un desbalance entre las clases que corresponden a los grados de Mercalli en el rango 1 al 7, por lo que aplicamos un remuestreo estratificado para generar uniformidad en la cantidad de instancias por clase. En la evaluación del modelo, usamos principalmente el error promedio absoluto (MAE), el coeficiente de correlación y la raíz de la desviación cuadrática media (RMSD).

Por último, los experimentos expuestos anteriormente fueron ejecutados con cada conjunto de métricas definidas, las que corresponden a las léxicas, de red y la combinación de ambas.

5.4. Aplicación del estimador espacial

El modelo generado anteriormente para predecir la intensidad de Mercalli en una zona afectada por un evento sísmico, no considera la dimensión geográfica. Además, las métricas léxicas y de red, no abarcan explícitamente la relación entre las comunas cercanas. Por lo que, el estimador espacial busca incorporar la dimensión geográfica en la predicción de Mercalli. Para lo anterior, aplicamos el método expuesto en la sección 4.3, sobre los resultados predictivos del modelo. La base de este método para incluir la dimensión espacial consiste en otorgarle relevancia a las comunas vecinas de la comuna objetivo.

Para definir la cantidad de vecinos relevantes[6] es importante considerar que si el número definido es muy grande, podría generar un mayor error en la estimación de intensidad por cada comuna objetivo, al ser afectadas por otras comunas extremadamente lejos de ellas. Notamos los efectos positivos para la estimación de Mercalli de una comuna, en los casos que incluyen la capital de la región como uno de los vecinos de dicha comuna. Explicamos este comportamiento, por la presencia en las capitales de una densa población, lo que garantiza

[5]https://www.cs.waikato.ac.nz/ ml/weka/index.html

[6]La cantidad de vecinos corresponde a la variable k, utilizado en el algoritmo k-vecinos cercanos.

una mayor posibilidad que personas informen debidamente en Twitter un evento sísmico que afecte dicha zona. Por lo anterior, definimos la regla de incluir dentro de los vecinos de cada comuna, la capital de la misma región.

Capítulo 6

Resultado y discusión

En este capítulo, exponemos los resultados generados de los experimentos detallados en el capítulo 5 y posteriormente realizamos el análisis, comparación y discusión de estos resultados. Los resultados están separados en *métricas*, *región de interés* y *estimador espacial*. La primera subsección presenta los resultados al analizar y comparar las métricas que son usadas posteriormente en la creación de los modelos. La siguiente subsección presenta los resultados obtenidos en la creación del modelo para predecir las comunas afectadas por un sismo. La tercera y última subsección muestra los resultados del modelo estimador de intensidad de Mercalli. En las dos últimas secciones, aplicamos cada experimento para 3 conjuntos de métricas, que corresponden a léxicas, de red y la unión de ambas.

6.1. Resultados

Para la experimentación, como mencionamos en la sección 5.1.2, utilizamos 331 sismos ocurridos en Chile entre el 2017 y 2018. El conjunto de sismos fue dividido en conjunto de entrenamiento y de prueba, 263 y 68 instancias respectivamente. En la Tabla 6.1, mostramos la distribución de los sismos según la intensidad máxima de Mercalli de cada evento, por otro lado, en la Tabla 6.2, mostramos la distribución de intensidad de Mercalli registrada a nivel de comunas; observamos en esta última tabla, un mayor número de registros debido a que los sismos generalmente son percibidos en una extensa área, lo cual, provoca registros de diversas intensidades de Mercalli en distintas comunas para un único evento sísmico. Destacamos en ambas tablas, la baja frecuencia de registros de intensidades altas en el periodo que comprende el conjunto de datos, esto lo explicamos mediante la Tabla 2.1, donde se muestran sismos con valores altos en escala Richter en baja cantidad; y si consideramos la relación proporcional existente entre ambas escalas (Richter y Mercalli), explicaría la baja frecuencia de intensidades mayores a 5 obtenidos en nuestros datos. Además advertimos la baja cantidad de registros con una intensidad de Mercalli de valor menor a 3, comparativamente con registros entre 4 y 5. Esto es explicado gracias a la Tabla 2.2, la cual muestra que en intensidades menores a 3, la capacidad de una persona en percibir un sismo, depende de condiciones especiales del ambiente, como ejemplo, encontrarse en altura al momento del

sismo.

Intensidad Máxima	2	3	4	5	6	7
Entrenamiento	11	105	103	39	3	2
Prueba	3	26	26	10	2	1
Total	14	131	129	49	5	3

Tabla 6.1: Distribución de entrenamiento/-prueba según la intensidad máxima de Mercalli por cada evento sísmico.

Intensidad	0	1	2	3	4	5	6	7
Entrenamiento	5021	837	1489	1921	893	282	38	10
Prueba	1527	196	319	413	223	70	95	4
Total	6548	1033	1808	2334	1116	352	133	14

Tabla 6.2: Distribución de entrenamiento/prueba según la intensidad registrada en cada comuna.

6.1.1. Métricas

Las métricas o características provenientes de Twitter, buscan caracterizar un evento sísmico, siendo estos datos empleados para generar el modelo de predicción de intensidad en escala Mercalli. En primera instancia, de forma gráfica, analizamos el comportamiento de cada métrica con respecto a la intensidad de una comuna en cierto evento sísmico. En última instancia, estudiamos el comportamiento de las métricas, buscando la existencia de una correlación entre ellas.

En la Figura 6.1, visualizamos un gráfico por cada una de las característica léxica con respecto a la escala Mercalli. En primer lugar, observamos la correlación existente entre el `NÚMERO DE TWEETS` y la escala Mercalli, donde a mayor cantidad de tweets en una comuna, es más probable que se presente un evento con alta intensidad. De igual forma, observamos un similar comportamiento para `TWEETS NORMALIZADO`, diferenciándose por la varianza observada, la cual, es menor para el caso de los `TWEETS NORMALIZADO`. En segundo lugar, observamos en el `PROMEDIO DE PALABRAS` y `LARGO PROMEDIO`, un decrecimiento de la mediana, y a su vez, presentan una estabilidad del uso de palabras y letras en escalas altas, lo que refuerza trabajos anteriores que ilustran que a mayor envergadura del evento sísmico las personas expresan la situación en mensajes más cortos. En tercer lugar, el `PROMEDIO DE SIGNOS DE INTERROGACIÓN` y `PROMEDIO DE SIGNOS DE EXCLAMACIÓN`, presentan un crecimiento de la mediana a medida que aumenta la intensidad del sismo. En cuarto lugar, el `USO DE MAYÚSCULAS`, observamos una baja presencia en el conjunto de datos, a excepción de sismos de escalas altas, por lo que podría considerarse una variable no significativa. En quinto lugar, el mayor `USO DE HASHTAG` esta presente en eventos con alta intensidad de Mercalli. Por otra parte, `USO DE RETWEETS` y `USO DE MENTION` presentan el efecto contrario, disminuyen a medida que aumenta el sismo. En último lugar, el gráfico de `POBLACION` observamos un de-

crecimiento de la mediana a medida que aumenta la intensidad de Mercalli. Por el contrario, para el uso de la palabra *terremoto*, observamos un aumento para eventos de alta intensidad.

En la Figura 6.2, exponemos los gráficos por cada característica de red con respecto a la escala Mercalli. Observamos en los gráficos que las características presentan una correlación lineal con respecto a la intensidad de un sismo a excepción del `Nº DE SEMILLAS S/ SOLITARIOS`, donde no muestra un comportamiento claramente definido. Además en algunas gráficos presentan un comportamiento errático para sismos de grado 2 en escala Mercalli, al ser una intensidad de baja magnitud, no todas las personas perciben el sismo, lo que genera mayor ruido en estos datos. En otro aspecto, destacamos lo observado en el gráfico del `PROMEDIO DE GRADOS`, al presentar una baja varianza en todos sus niveles de Mercalli.

En la Tabla 6.3, mostramos el coeficiente de correlación calculada con el método de Spearman, entre las métricas léxicas. En el primer bloque de dos filas mostramos la correlación de la primera métrica, `MERCALLI`, con cada una de las otras doce métrica. En el segundo bloque de dos filas, mostramos la correlación de la segunda métrica, `NUMERO DE TWEETS`, con cada una de las métrica restantes (exceptuando `MERCALLI`), y así en el resto de la tabla. Observamos una positiva correlación entre el `NUMERO DE TWEETS` y `TWEETS NORMALIZADOS`. Por otro lado, existe una negativa correlación entre `TWEETS NORMALIZADOS` y `POBLACION`. Como esperabamos existe una correlación entre de dos características provenientes del largo de un tweet (`PROMEDIO DE PALABRAS` y `LARGO PROMEDIO`), y también entre `SÍMBOLO MENTION` y `SÍMBOLO RETWEET`, considerando que este último es un subconjunto del anterior. Esto ocurre porque los mensajes que son retweet siempre incluyen un símbolo de mención del autor del mensaje original.

	MERCALLI	NUMERO DE TWEETS	TWEETS NORMALIZADOS	PROMEDIO DE PALABRAS	LARGO PROMEDIO	SIGNO DE INTERROGACIÓN	SIGNO DE EXCLAMACIÓN	PALABRAS EN MAYÚSCULAS	SÍMBOLO HASHTAG	SÍMBOLO MENTION	SÍMBOLO RETWEET	PALABRA "TERREMOTO"	POBLACIÓN
MERCALLI	ρ	0.22	0.24	-0.11	-0.14	0.05	0.16	0.15	0.06	-0.12	-0.11	0.05	-0.02
	p	2.2e-16	2.2e-16	2.2e-16	2.2e-16	3.3e-13	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16	6.2e-12	0.0
NUMERO DE TWEETS	ρ	**0.52**	-0.09	-0.18	0.29	0.38	0.22	0.21	-0.22	-0.17	0.25	0.3	
	p	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16	
TWEETS NORMALIZADOS	ρ	-0.07	-0.16	0.19	0.29	0.18	0.12	-0.14	-0.11	0.14	**-0.56**		
	p	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16		
PROMEDIO DE PALABRAS	ρ	**0.86**	0.02	-0.11	-0.08	-0.04	0.31	0.33	0.02	0.0			
	p	2.2e-16	9.5e-05	2.2e-16	2.2e-16	3.1e-10	2.2e-16	2.2e-16	0.002	0.201			
LARGO PROMEDIO	ρ	-0.08	-0.18	-0.1	-0.04	**0.41**	**0.43**	0.0	0.0				
	p	2.2e-16	2.2e-16	2.2e-16	1.1e-09	2.2e-16	2.2e-16	0.249	0.956				
SIGNO DE INTERROGACIÓN	ρ	0.23	0.08	0.13	-0.04	-0.01	0.1	0.06					
	p	2.2e-16	2.2e-16	2.2e-16	3.1e-10	0.031	2.2e-16	2.2e-16					
SIGNO DE EXCLAMACIÓN	ρ	0.18	0.11	-0.14	-0.12	0.23	0.09						
	p	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16						
PALABRAS EN MAYÚSCULAS	ρ	0.04	-0.07	-0.05	0.12	0.09							
	p	6.7e-10	2.2e-16	4.1e-11	2.2e-16	2.2e-16							
SÍMBOLO HASHTAG	ρ	-0.13	-0.07	0.07	0.06								
	p	2.2e-16	2.2e-16	2.2e-16	2.2e-16								
SÍMBOLO MENTION	ρ	**0.86**	-0.07	-0.07									
	p	2.2e-16	2.2e-16	2.2e-16									
SÍMBOLO RETWEET	ρ	-0.07	-0.04										
	p	2.2e-16	9.8e-09										
PALABRA "TERREMOTO"	ρ	0.15											
	p	2.2e-16											

Tabla 6.3: Coeficiente de Correlación de Spearman de las características léxicas consideradas en nuestro estudio. Los coeficiente de Spearman encontrados son estadísticamente significativos como muestra el p-value. Las correlaciones fuertes son indicados en negrita.

En la Tabla 6.4, exponemos el coeficiente de correlación calculada con el método de Spearman, entre las métricas de red. Observamos la correlación de MERCALLI con PROMEDIO NODOS, PROMEDIO NODOS S/SOLITARIOS, PROMEDIO GRADOS, PROMEDIO GRADOS S/SOLITARIOS Y MAX. GRADOS. Además destacamos la fuerte correlación existente entre las métricas relacionadas con nodos y grados (PROMEDIO NODOS, PROMEDIO NODOS S/SOLITARIOS, MAX. NODOS, PROMEDIO GRADOS , PROMEDIO GRADOS S/SOLITARIOS Y MAX. GRADOS). En planos generales, las métricas de red presentan un alto nivel de correlación entre ellas y a su vez, con MERCALLI.

	MERCALLI	N DE SEMILLAS	N DE SEMILLAS S/SOLITARIOS	PROMEDIO NODOS	PROMEDIO NODOS S/SOLITARIOS	MAX. NODOS	PROMEDIO GRADOS	PROMEDIO GRADOS S/SOLITARIOS	MAX. GRADOS	POBLACIÓN
MERCALLI	ρ	0.15	0.18	0.21	0.21	0.20	0.21	0.21	0.21	-0.15
	p	2.2e-16	2.2e-16	2.2e-16	2.2e-16	3.3e-13	2.2e-16	2.2e-16	2.2e-16	2.2e-16
N DE SEMILLAS	ρ		0.55	0.50	0.56	0.56	0.51	0.55	0.56	0.48
	p		2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16
N DE SEMILLAS S/SOLITARIOS	ρ			0.96	0.97	0.97	0.96	0.97	0.97	0.34
	p			2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16
PROMEDIO NODOS	ρ				0.99	0.98	0.99	0.99	0.99	0.31
	p				2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16
PROMEDIO NODOS S/SOLITARIOS	ρ					0.99	0.99	0.99	0.99	0.35
	p					2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16
MAX. NODOS	ρ						0.99	0.99	0.99	0.35
	p						2.2e-16	2.2e-16	2.2e-16	2.2e-16
PROMEDIO GRADOS	ρ							0.99	0.99	0.32
	p							2.2e-16	2.2e-16	2.2e-16
PROMEDIO GRADOS S/SOLITARIOS	ρ								0.99	0.34
	p								2.2e-16	2.2e-16
MAX. GRADOS	ρ									0.35
	p									2.2e-16

Tabla 6.4: Coeficiente de Correlación de Spearman de las características de red consideradas en nuestro estudio. Los coeficiente de Spearman encontrados son estadísticamente significativos como muestra el p-value.

6.1.2. Región de interés

Construimos el primer modelo, para predecir las comunas que percibieron un evento sísmico, denominadas zonas de interés. Realizamos 3 experimentos, diferenciados por el conjunto de métricas, las léxicas, de red y la combinación de ambas. En la creación del modelo, probamos distintos algoritmos como Naive Bayes o Multilayer Perceptron, pero el mejor resultado obtenido en los 3 experimentos fue utilizando el algoritmo SVM. Los resultados de los otros algoritmos, se encuentran en el apéndice A.

Al aplicar las métricas léxicas usando el modelo basado en SVM, obtuvimos 7325 instancias clasificadas correctamente, que representan el 69.82 %. En las Tablas 6.5 y 6.6 mostramos el detalle de los resultados del modelo sobre el conjunto de entrenamiento y de prueba, respectivamente.

Class	FP Rate	Precision	Recall	F-measure	ROC Area
0	0.189	0.736	0.575	0.646	0.693
1	0.425	0.675	**0.811**	0.737	0.693
W. Avg.	0.312	0.705	0.698	0.693	0.693

Tabla 6.5: Métricas léxicas: resultados del conjunto de entrenamiento por clase de la *región de interés*, usando 5-fold cross validation.

Al aplicar el modelo generado con el conjunto de entrenamiento sobre los datos de prueba, obtenemos 1867 instancias clasificadas correctamente de un total de 2847, logrando un 65.56 % de acierto. Esto nos indica que el clasificador es lo bastante generalizado, para aplicar en distintos grupos de datos. En el caso de la precisión, presenta un bajo nivel, lo que se explica por la existencia de ruido en los datos a nivel de comuna. A pesar de esto, hacemos notar el alto *Recall* en la clase 1, indicando la predictibilidad sobre la clase 1 (zonas que perciben el sismo). En resumen, el modelo sobrestima la clase 1, abarcando en gran medida las regiones de interés.

Class	FP Rate	Precision	Recall	F-measure	ROC Area
0	0.184	0.765	0.517	0.617	0.666
1	0.483	0.593	**0.816**	0.687	0.666
W. Avg.	0.323	0.685	0.655	0.649	0.666

Tabla 6.6: Métricas léxicas: resultados del conjunto de prueba por clase de la *región de interés*, usando 5-fold cross validation.

La métricas de red, lograron en el modelo generado 7831 instancias correctas, que equivale al 74.64 % del conjunto de datos de entrenamiento. Para el conjunto de prueba fueron 1968 instancias correctas abarcando 69.12 %. El detalle de los resultados lo observamos en las Tablas 6.7 y 6.8.

Observamos 3 aspectos relevantes en los resultados expuestos. En primer lugar, el modelo aplicado en el conjunto de entrenamiento obtuvo un *Recall* 0.805, lo cual nos indica un

Class	FP Rate	Precision	Recall	F-measure	ROC Area
0	0.195	0.763	0.683	0.721	0.744
1	0.317	0.734	**0.805**	0.768	0.744
W. Avg.	0.259	0.748	0.746	0.745	**0.744**

Tabla 6.7: Métricas de red: resultados del conjunto de entrenamiento por clase de la *región de interés*, usando 5-fold cross validation.

Class	FP Rate	Precision	Recall	F-measure	ROC Area
0	0.207	0.771	0.603	0.677	0.698
1	0.397	0.633	**0.793**	0.704	0.698
W. Avg.	0.295	0.707	0.691	0.690	**0.698**

Tabla 6.8: Métricas de red: resultados del conjunto de prueba por clase de la *región de interés*, usando 5-fold cross validation.

alto acierto de la clase 1 respecto al universo de la clase. Además este comportamiento es observado en el conjunto de prueba con un *Recall* de 0.793, menor al de entrenamiento, pero con una diferencia poco significativa. En segundo lugar, el uso de métricas de red logró un *ROC Area* de 0.744, lo que nos indica un buen rendimiendo del clasificador, en comparación con las métricas léxicas. En tercer y último lugar, el *F-measure*, nos indica un buen modelo contemplando la precisión y *Recall*.

Combinamos ambas métricas como un único conjunto de métricas. Utilizando los datos de entrenamiento, el modelo mostró 7416 instancias correctas que equivalen al 70.68 %. En el caso de los datos de prueba, fueron 1874 correctas que sería 65.82 %. El detalle de los resultados del modelo lo observamos en la Tabla 6.9, donde destacamos el alto *Recall* de 0.813, que nos indica en el modelo, una correcta clasificación de la clase 1 (comunas que perciben el sismo). Observamos en la Tabla 6.10, los resultados del modelo sobre el conjunto de prueba. Al igual que los datos de entrenamiento, obtuvimos un alto indicador de *Recall*, y más aún, logramos un mayor valor que los casos de entrenamiento, de 0.820.

Class	FP Rate	Precision	Recall	F-measure	ROC Area
0	0.187	0.744	0.591	0.659	0.702
1	0.409	0.684	**0.813**	0.743	0.702
W. Avg.	0.303	0.713	0.707	0.703	0.702

Tabla 6.9: Métricas léxicas y red: resultados del conjunto de entrenamiento por clase de la *región de interés*, usando 5-fold cross validation.

Para comprender mejor el comportamiento de nuestro clasificador 0/1 de región de interés, exponemos el detalle del resultado del modelo para cada conjunto de métricas de forma desagregada, es decir, el detalle de la clasificación por cada nivel de la escala Mercalli. En las Tablas 6.11, 6.12 y 6.13 observamos para altas intensidades (escala Mercalli igual o superior a 5), una baja tasa de error o incluso nula en la predicción de comunas que perciben el sismo. Este aspecto es muy importante, ya que muestra el buen funcionamiento del modelo sobre

Class	FP Rate	Precision	Recall	F-measure	ROC Area
0	0.180	0.769	0.519	0.619	0.669
1	0.481	0.595	**0.820**	0.690	0.669
W. Avg.	0.320	0.689	0.658	0.652	0.669

Tabla 6.10: Métricas léxicas y red: resultados del conjunto de prueba por clase de la *región de interés*, usando 5-fold cross validation.

zonas impactadas por una intensidad de Mercalli elevada. Además en las métricas de red, se observa una menor tasa de error en la predicción de las zonas que no perciben sismos, en comparación con los otros conjuntos de métricas.

Act.	Pred.	-	1	2	3	4	5	6	7
0	0	790	-	-	-	-	-	-	-
0	1	737	-	-	-	-	-	-	-
1	0	-	66	85	62	25	5	-	-
1	1	-	130	234	351	198	65	95	4
Instancias		1527	196	319	413	223	70	95	4
Tasa de error		0.48	0.33	0.26	0.15	0.11	0.07	-	-

Tabla 6.11: Desagregación métricas léxicas: instancias de prueba divididas por intensidades de Mercalli y el valor real y predecido de la región de interés.

Act.	Pred.	-	1	2	3	4	5	6	7
0	0	921	-	-	-	-	-	-	-
0	1	606	-	-	-	-	-	-	-
1	0	-	83	85	65	40	-	-	-
1	1	-	113	234	348	183	70	95	4
Instancias		1527	196	319	413	223	70	95	4
Tasa de error		0.39	0.42	0.26	0.15	0.17	-	-	-

Tabla 6.12: Desagregación métricas de red: instancias de prueba divididas por intensidades de Mercalli y el valor real y predecido de la región de interés.

6.1.3. Estimador espacial

Como explicamos en las secciones 4.2 y 4.3 utilizamos una regresión y suavizador espacial para estimar la intensidad de Mercalli. En el caso de la regresión, usamos el modelo de regresión de vector de soporte con una optimización mínima secuencial (SMO). En la Tabla 6.14 se muestran los resultados del modelo sobre el algoritmo SMO para los 3 conjuntos de métricas (léxicas, de red y combinadas).

Observamos, que al evaluar el modelo sobre los datos de prueba, en los 3 conjuntos de métricas, empeora considerablemente los criterios de evaluación. Este resultado nos indica,

Act.	Pred.	-	1	2	3	4	5	6	7
0	0	792	-	-	-	-	-	-	-
0	1	735	-	-	-	-	-	-	-
1	0	-	71	88	50	25	4	-	-
1	1	-	125	231	363	198	66	95	4
Instancias		1527	196	319	413	223	70	95	4
Tasa de error		0.48	0.36	0.27	0.12	0.11	0.05	-	-

Tabla 6.13: Desagregación métricas léxicas y red: instancias de prueba divididas por intensidades de Mercalli y el valor real y predecido de la región de interés.

Conjunto	**Evaluación**	**Léxicas**	**Red**	**Léxicas y Red**
Entrenamiento	Coef. de correlación	0.65	0.68	0.67
	MAE	1.15	1.13	1.12
	RMSD	1.54	1.47	1.51
Prueba	Coef. de correlación	0.26	0.42	0.28
	MAE	2.26	1.91	2.10
	RMSD	2.83	2.33	2.68

Tabla 6.14: Resultado del modelo SMO. Primera columna indica si fue aplicado el modelo sobre el conjunto de entrenamiento o prueba. A continuación indica el tipo de evaluación aplicado en el modelo. Finalmente las últimas 3 columnas indican las métricas empleadas en el modelo, léxicas, de red o combinación de ambas.

que el modelo de regresión es insuficiente para lograr una correcta predicción. Debido a lo anterior, usamos sobre el modelo generado, nuestro ajuste de Mercalli y método de suavizado espacial, definidos en secciones 4.3. En el método, evaluamos el MAE de diferentes valores de λ entre el real Mercalli y el estimado. Por cada sismo del conjunto de prueba, promediamos el MAE de las comunas y generamos un único MAE por sismo. Entonces, estratificamos el MAE según la máxima intensidad de Mercalli por sismo y calculamos el MAE por nivel de Mercalli. Por ejemplo, MAE(5) corresponde al promedio de MAE de todos los sismos del conjunto de prueba con intensidad máxima 5 en la escala de Mercalli. Como la distribución de los sismos por nivel de Mercalli es desbalanceada, utilizamos el *MAE general*, definido en la ecuación 6.1, donde castigamos el error proporcional a la intensidad del sismo.

$$\text{MAE General} = \frac{\sum_{M \in \text{escala Mercalli}} \text{MAE}(M) \cdot M \cdot \#\text{Instancias de M}}{\sum_{M \in \text{escala Mercalli}} M \cdot \#\text{Instancias de M}} \tag{6.1}$$

En la Tabla 6.15, mostramos el *MAE General* para distintos valores de λ $\{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$, de los modelos generados con los 3 conjuntos de métricas. Observamos una mejora del valor obtenido del *MAE General*, a medida que aumenta el valor de λ, hasta llegar a 0.8, siendo este comportamiento visible en el resultado de los 3 experimentos. Por otro lado, comparando los resultados entre métricas, es posible destacar el modelo de métricas léxicas, ya que presenta un menor *MAE General*, lo que nos indica un menor error para la estimación de intensidad de Mercalli.

λ	0	0.2	0.4	0.6	0.8	1.0
Métricas Léxicas	2.07	1.84	1.55	1.20	**0.87**	1.02
Métricas de Red	2.29	2.13	1.37	1.27	**1.21**	1.34
Métricas Léxica y de Red	2.20	2.24	1.62	1.35	**1.18**	1.29

Tabla 6.15: MAE General para diferentes valores de λ en los 3 conjuntos de métricas usados en los experimentos.

Finalmente, escogemos el modelo proveniente de las métricas léxicas y aplicando el método de estimador espacial sobre este, con un λ igual a 0.8; mostramos un ejemplo ilustrativo del estimador espacial de Mercalli, en el caso de un sismo con intensidad máxima 7 en escala Mercalli. Podemos comparar en la Tabla 6.16, el valor estimado de intensidad en comunas afectadas por el sismo, respecto al valor de intensidad reportado.

M	Reporte Real	Reporte Estimado
3	Talca (3), Constitución (3), Maule (3), Navidad (2), Santiago (1)	Maule (4), Talca (3), Constitución (3), Curepto (3)
4	Coquimbo (4), Ovalle (4), Melipilla (2), Santiago (2)	Coquimbo (3), Colina (3), Til-Til (3), Santiago (3)
5	La Serena (5), Coquimbo (5), Vicuña (4), Ovalle (4), Illapel (3)	La Serena (5), Coquimbo (5), Vicuña (4), Ovalle (4), Illapel (3)
6	Quintero (6), Valparaíso (5), Quilpue (5), Quillota (5), Viña del Mar (4), Ovalle(3), Santiago(3)	Quintero (6), Valparaíso (6), Viña del Mar (6), Quilpue (5), Quillota (5), San Felipe (4), Santiago (4), La Serena (3)
7	Limache(7), Santiago (6), Viña del Mar (6), Valparaíso (6), Coquimbo (5), La Serena (5), Ovalle (5), Rancagua (5), Curicó (4), Coronel (3)	Limache (6), Viña del Mar (6), Valparaíso (6), Santiago (5), Coquimbo (5), La Serena (5), Rancagua (5), Curicó (5), Ovalle (4), Quirihue (3)

Tabla 6.16: Ejemplo comparativo del registro real y estimado de intensidad de Mercalli en un evento sísmico ocurrido en Chile. Cada fila es un ejemplo de evento sísmico, donde la primera columna indica la intensidad de Mercalli máximo de dicho evento.

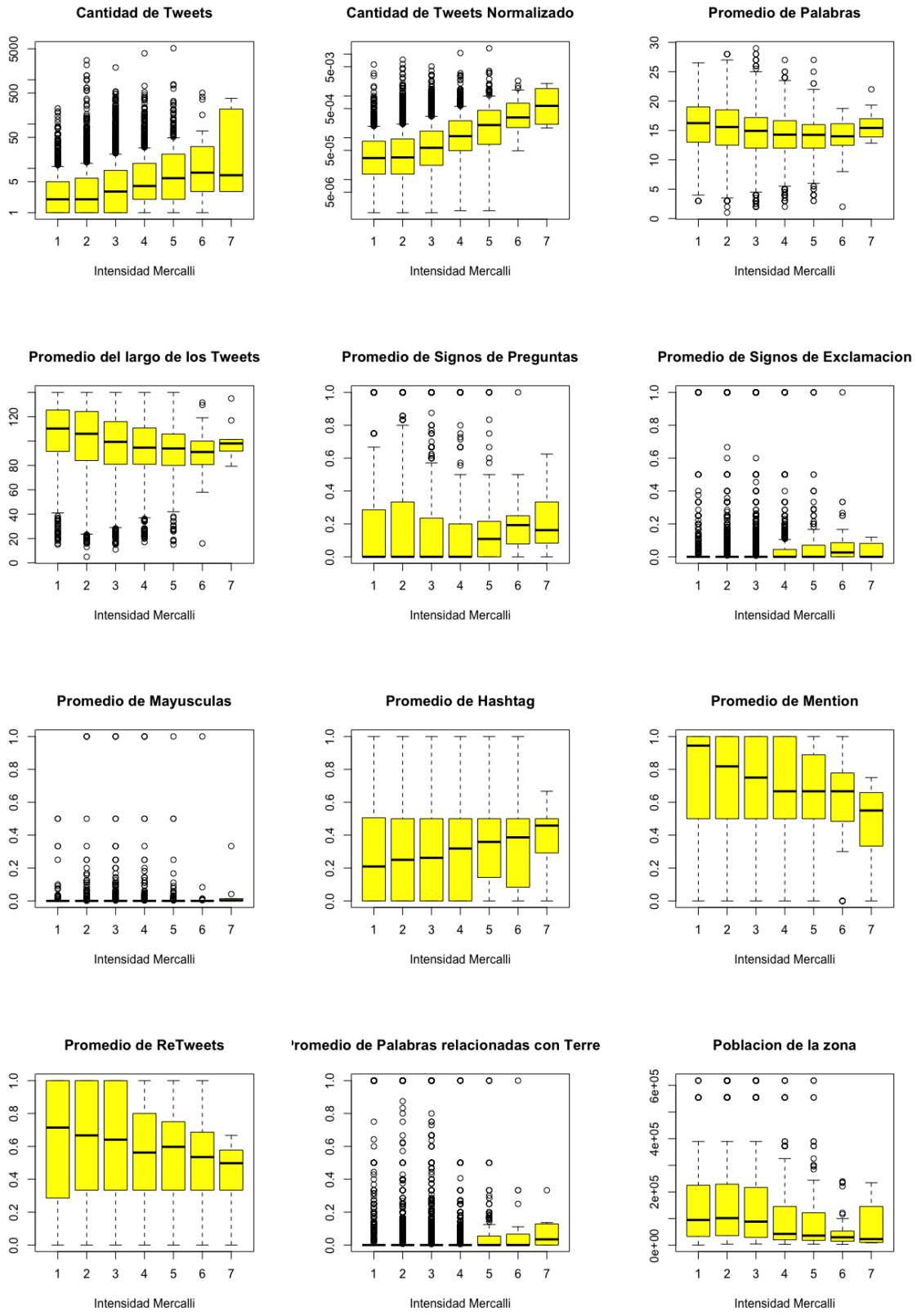


Figura 6.1: Gráficos de caja por cada métrica léxica. En algunos gráficos, aplicamos la escala logarítmica, para mejorar la visualización de los datos.

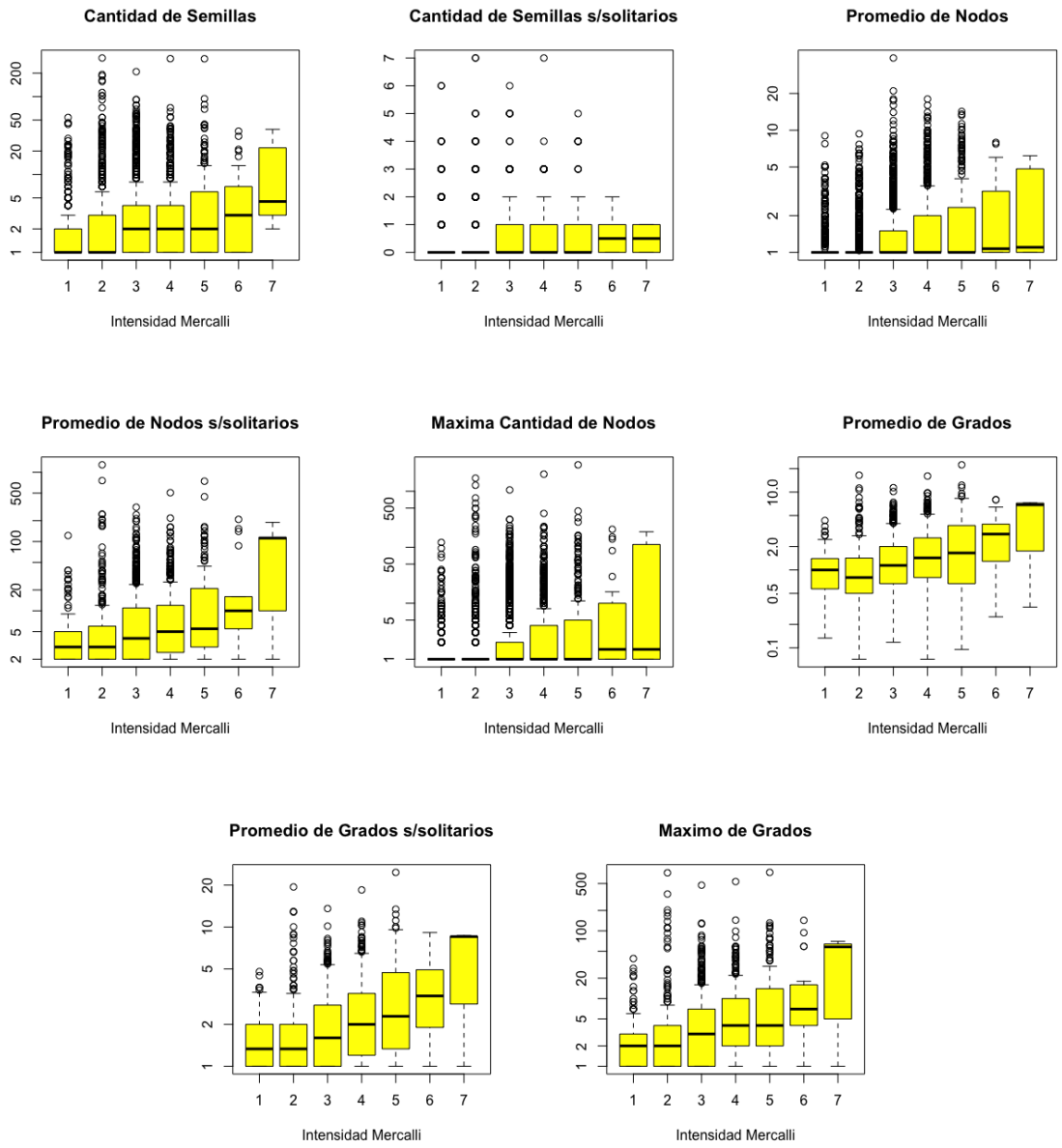


Figura 6.2: Gráficos de caja por cada métrica de red. En algunos gráficos, aplicamos la escala logarítmica, para mejorar la visualización de los datos.

6.2. Discusión

Los resultados experimentales, expuestos en la sección anterior, nos han mostrado el potencial de las métricas desarrolladas (léxicas y de red) para caracterizar un evento sísmico.

Un tema relevante en este trabajo, ha sido la dimensión geográfica del fenómeno extraída de Twitter. Destacamos este aspecto, ya que permite caracterizar el comportamiento de una localidad, frente a un evento sísmico. A pesar del bajo uso de las coordenadas geográficas por parte de los usuarios, ha sido posible inferir dicha información a partir de otras fuentes relacionadas con la geografía, como el campo *localización* del tweet y del usuario. Este proceso fue fundamental en el trabajo, ya que permitió generar las métricas de red, las cuales por primera vez se han utilizado en trabajos relacionados con sismos.

Se debe considerar que en un evento sísmico las personas transmiten lo sucedido a través de redes sociales, tanto por ser afectadas o con fines de difundir o informarse de lo ocurrido. Debido a esto, el primer experimento buscó diferenciar las comunas que percibieron el sismo y las que no. Destacamos que los 3 conjuntos de experimentos (métricas léxicas, de red y combinación de las anteriores) lograron un alto porcentaje para identificar las zonas que realmente son afectadas. Entre los 3 conjuntos, mencionamos las métricas de red, al presentar el mejor valor de *ROC Area*, indicando de forma positiva el modelo generado a partir de estas métricas. Para entender mejor los modelos que predicen las zonas de interés, desagregamos los resultados según el Mercalli de la zona y notamos una mayor precisión en magnitudes altas en escala Mercalli. Esto indica que las métricas utilizadas pueden caracterizar sismos, especialmente en zonas que presentan una mayor intensidad Mercalli. Aplicar los modelos generados para los distintos conjuntos de datos, permite verificar la generalización de los modelos, al ser aplicados sobre conjuntos distintos de datos. En nuestro estudio, los modelos presentan buenos resultados, tanto para el conjunto de entrenamiento como de prueba.

A partir de los buenos resultados para predecir la región de interés, proseguimos con la creación de un modelo para estimar la intensidad de Mercalli a nivel de comuna. Inicialmente el problema es considerado discreto, al tener que predecir valores enteros entre 1 al 8 (escala de Mercalli). Pero si comprendemos la intensidad, como un efecto continua en las zonas contiguas, hace sentido abarcar el problema como un regresión, por lo que generamos los modelos de los distintos conjuntos de métricas, con el método SMO. En los tres conjuntos, se presentan similares resultados para los parámetros de evaluación, Coef. de correlación, MAE y RMSD. Pero si observamos los resultados al aplicar el modelo sobre el conjunto de prueba, baja la calidad de precisión. A partir de este problema, introducimos la dimensión espacial para caracterizar de mejor manera el evento sísmico por comuna.

Los resultados de la Tabla 6.15 expusieron que al aplicar el estimador espacial, observamos una disminución del error de la estimación, a medida que damos mayor relevancia a las zonas vecinas de la comuna objetivo. Esto lo reflejamos al definir el λ óptimo en 0.8, el cual se interpreta en la relación 4 es a 1, donde la mayor relevancia para estimar la intensidad Mercalli de una comuna recae en las estimaciones de las comunas vecinas. Finalmente al comparar los modelos de cada conjunto de métricas observamos la predominancia de las características léxicas al momento de disminuir el error en la estimación.

Capítulo 7

Conclusiones

En este trabajo hemos presentado cómo encontrar la relación entre eventos sísmicos ocurridos en Chile y la información generada en la red social Twitter. Este estudio, incluyó la definición de las métricas de ambas áreas, proponer una metodología a aplicar, realizar los experimentos a partir de la metodología y finalmente exponer los resultados junto a la discusión de estos. En este capítulo, discutimos las principales contribuciones y hallazgos de este trabajo que aportan al conocimiento.

7.1. Contribución y relevancia

En este trabajo se presentaron importantes resultados que además de entregar nuevo conocimiento permiten avalar investigaciones anteriores relacionadas con redes sociales y sismos. A continuación exponemos las contribuciones más relevantes.

En esta investigación, consideramos distintos tipos de características proveniente de Twitter. Al usar las métricas léxicas, hemos logrado revalidar investigaciones anteriores con resultados similares en la relación de estas métricas a la caracterización del evento sísmico aplicado en Chile. Además cabe mencionar que el uso de métricas de red no se encuentra en la literatura, por lo que sería la primera vez que se emplean para generar modelos predictivos en sismos, teniendo igual o mejores resultados que las métricas léxicas.

La metodología definida en este trabajo, para lograr la correlación entre métricas de Twitter y eventos sísmicos, puede ser aplicada para otro conjunto de datos. El requisito más importante del conjunto de datos, es lograr un alto porcentaje de recuperación de la geolocalización de los tweets, con el objeto de aplicar el método de estimador espacial en la etapa de generación de modelo.

Logramos plantear el método de estimación espacial de Mercalli, cuyo objetivo es mejorar la predicción del modelo para la estimación de Mercalli a partir de métricas originadas en Twitter. Este método se basó en estimar el Mercalli de una localidad a partir de la información en las localidades cercanas dando relevancia a la dimensión del espacio para estimar el

Mercalli. Los resultados expusieron una mejora significativa al aplicar el método destacando que es primera vez que aplica un método espacial para modelos predictivos de sismos.

Destacamos la caracterización lograda de los eventos sísmicos, con el modelo generado. Esto es reflejado, al obtener alta precisión en sismos de alta intensidad, dando posible explicación de la imprecisión presente en eventos sísmicos con un grado de Mercalli bajo.

Twitter ha sido una fuente importante de investigación para la caracterización de eventos físicos, en particular, eventos sísmicos. Actualmente presenta innumerables desafíos para las ciencias de la computación, incluso con este trabajo, se abren nuevas ramas que son expuestas en la sección 7.2.

7.2. Trabajo futuro

El análisis entre métricas de Twitter y sismos que se expuso en este trabajo, permite abrir nuevas líneas de investigación. En esta sección se muestran y discuten algunas de ellas como potenciales trabajos.

Primero que todo, considerando que los experimentos fueron con datos de sismos ocurridos en Chile, es relevante verificar la efectividad de la metodología empleada replicándola en otro conjunto de datos. Como se utilizaron eventos situados en Chile, sería un gran aporte, emplear la metodología con datos de otro país. Esto permitirá contrastar el comportamiento de un país con otro, analizando tanto las similitudes y diferencias presentes con la misma metodología. Sin embargo, una dificultad identificable debido al trabajo previo, se vincula con la obtención de datos de Twitter que se asocian a un evento sísmico. Esto se produce por la presencia de ruido en los datos e información incompleta, como por ejemplo, utilizar una palabra que se asocia a sismo, en un evento no sísmico o no incluir información de geolocalización, complicando la generación de las variables de los experimentos.

Por otro lado, un tema muy relevante, fueron las métricas que se utilizaron, las cuales, fueron divididas en dos categorías: léxicas y de red. Para nuevas investigaciones, sería relevante el incluir nuevas métricas, y así comparar los resultados y evaluar cuáles describen mejor el evento sísmico. A partir de esta idea, uno de los enfoques interesantes a investigar, es la semántica de los tweets, donde se puede evaluar métricas relacionadas con sentimientos o emociones.

Considerando los buenos resultados obtenidos al aplicar la estimación espacial en este trabajo, sería prometedor su aplicación en otras área de estudio, siempre que incluyan la dimensión geográfica como una variable de investigación.

Por último, en relación a lo expuesto en este trabajo para solucionar la dificultad de recopilar información geolocalizada, se logró abarcar parcialmente este problema, lo cual, nos plantea el desafió de buscar otras alternativas que nos ayuden a lograr un mayor volumen de datos geolocalizados, para su uso en futuros experimentos.

Bibliografía

[1] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684. ACM, 2011.

[2] Makoto Okazaki and Yutaka Matsuo. Semantic twitter: Analyzing tweets for real-time event notification. In *Recent Trends and Developments in Social Software*, pages 63–74. Springer, 2010.

[3] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

[4] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Tweet analysis for real-time event detection and earthquake reporting system development. *IEEE Transactions on Knowledge and Data Engineering*, 25(4):919–931, 2013.

[5] Paul Earle, Michelle Guy, Richard Buckmaster, Chris Ostrum, Scott Horvath, and Amy Vaughan. Omg earthquake! can twitter improve earthquake response? *Seismological Research Letters*, 81(2):246–251, 2010.

[6] L Burks, M Miller, and R Zadeh. Rapid estimate of ground shaking intensity by combining simple earthquake characteristics with tweets. In *10th US Nat. Conf. Earthquake Eng., Front. Earthquake Eng., Anchorage, AK, USA, Jul. 21Y25*, 2014.

[7] Yury Kryvasheyeu, Haohui Chen, Nick Obradovich, Esteban Moro, Pascal Van Hentenryck, James Fowler, and Manuel Cebrian. Rapid assessment of disaster damage using social media activity. *Science advances*, 2(3):e1500779, 2016.

[8] Marco Avvenuti, Stefano Cresci, Andrea Marchetti, Carlo Meletti, and Maurizio Tesconi. Ears (earthquake alert and report system): a real time decision support system for earthquake crisis management. In *Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1749–1758. ACM, 2014.

[9] Bella Robinson, Robert Power, and Mark Cameron. A sensitive twitter earthquake detector. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 999–1002. ACM, 2013.

[10] Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012.

[11] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[12] Günther Sagl, Bernd Resch, Bartosz Hawelka, and Euro Beinat. From social sensor data to collective human behaviour patterns: Analysing and visualising spatio-temporal dynamics in urban environments. In *Proceedings of the GI-Forum*, pages 54–63, 2012.

[13] Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. Twitter under crisis: Can we trust what we rt? In *Proceedings of the first workshop on social media analytics*, pages 71–79. ACM, 2010.

[14] Stuart E Middleton, Lee Middleton, and Stefano Modafferi. Real-time crisis mapping of natural disasters using social media. *IEEE Intelligent Systems*, 29(2):9–17, 2014.

[15] Amanda Lee Hughes and Leysia Palen. Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management*, 6(3-4):248–260, 2009.

[16] James P Bagrow, Dashun Wang, and Albert-Laszlo Barabasi. Collective response of human populations to large-scale emergencies. *PloS one*, 6(3):e17680, 2011.

[17] Bertrand De Longueville, Robin S Smith, and Gianluca Luraschi. Omg, from here, i can see the flames!: a use case of mining location based social networks to acquire spatio-temporal data on forest fires. In *Proceedings of the 2009 international workshop on location based social networks*, pages 73–80. ACM, 2009.

[18] Tridib Mukherjee, Deepthi Chander, Sharanya Eswaran, Mridula Singh, Preethy Varma, Amandeep Chugh, and Koustuv Dasgupta. Janayuja: A people-centric platform to generate reliable and actionable insights for civic agencies. In *Proceedings of the 2015 Annual Symposium on Computing for Development*, pages 137–145. ACM, 2015.

[19] Konstantin Bauman, Alexander Tuzhilin, and Ryan Zaczynski. Using social sensors for detecting emergency events: a case of power outages in the electrical utility industry. *ACM Transactions on Management Information Systems (TMIS)*, 8(2-3):7, 2017.

[20] Rui Li, Kin Hou Lei, Ravi Khadiwala, and Kevin Chen-Chuan Chang. Tedas: A twitter-based event detection and analysis system. In *Data engineering (icde), 2012 ieee 28th international conference on*, pages 1273–1276. IEEE, 2012.

[21] Jheser Guzman and Barbara Poblete. On-line relevant anomaly detection in the twitter stream: an efficient bursty keyword detection model. In *Proceedings of the acm sigkdd workshop on outlier detection and description*, pages 31–39. ACM, 2013.

[22] Xin Zhang and Dennis Shasha. Better burst detection. In *Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on*, pages 146–146. IEEE, 2006.

[23] Michael Mathioudakis and Nick Koudas. Twittermonitor: trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pages 1155–1158. ACM, 2010.

[24] Luca D'Auria and Vincenzo Convertito. Real-time mapping of earthquake perception areas in the italian region from twitter streams analysis. In *Earthquakes and Their Impact on Society*, pages 619–630. Springer, 2016.

[25] Mark A Cameron, Robert Power, Bella Robinson, and Jie Yin. Emergency situation awareness from twitter for crisis management. In *Proceedings of the 21st International Conference on World Wide Web*, pages 695–698. ACM, 2012.

[26] Mor Naaman, Jeffrey Boase, and Chih-Hui Lai. Is it really about me?: message content in social awareness streams. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 189–192. ACM, 2010.

[27] Sasa Petrovic, Miles Osborne, and Victor Lavrenko. Rt to win! predicting message propagation in twitter. *ICWSM*, 11:586–589, 2011.

[28] Will Webberley, Stuart Allen, and Roger Whitaker. Retweeting: A study of message-forwarding in twitter. In *Mobile and Online Social Networks (MOSN), 2011 Workshop on*, pages 13–18. IEEE, 2011.

[29] Leticia Vidal, Gastón Ares, and Sara R Jaeger. Use of emoticon and emoji in tweets for food-related emotional expression. *Food Quality and Preference*, 49:119–128, 2016.

[30] Enrico Steiger, Joao Porto De Albuquerque, and Alexander Zipf. An advanced systematic literature review on spatiotemporal analyses of t witter data. *Transactions in GIS*, 19(6):809–834, 2015.

[31] Bumsuk Lee and Byung-Yeon Hwang. A study of the correlation between the spatial attributes on twitter. In *2012 IEEE 28th International Conference on Data Engineering Workshops*, pages 337–340. IEEE, 2012.

[32] Rodolfo Gonzalez, Gerardo Figueroa, and Yi-Shin Chen. Tweolocator: a non-intrusive geographical locator system for twitter. In *Proceedings of the 5th ACM SIGSPATIAL international workshop on location-based social networks*, pages 24–31. ACM, 2012.

[33] Clodoveu A Davis Jr, Gisele L Pappa, Diogo Rennó Rocha de Oliveira, and Filipe de L. Arcanjo. Inferring the location of twitter messages based on user relationships. *Transactions in GIS*, 15(6):735–751, 2011.

[34] Dennis Thom, Harald Bosch, Steffen Koch, Michael Wörner, and Thomas Ertl. Spatiotemporal anomaly detection through visual analysis of geolocated twitter messages. In *Visualization Symposium (PacificVis), 2012 IEEE Pacific*, pages 41–48. IEEE, 2012.

[35] Enrico Steiger, Bernd Resch, and Alexander Zipf. Exploration of spatiotemporal and semantic clusters of twitter data using unsupervised neural networks. *International Journal of Geographical Information Science*, 30(9):1694–1716, 2016.

[36] Barbara Poblete, Jheser Guzman, Jazmine Maldonado, and Felipe Tobar. Robust detection of extreme events using twitter: Worldwide earthquake monitoring. *IEEE Transactions on Multimedia*, 2018.

[37] Leonard Richardson. Beautiful soup documentation, 2007.

Apéndice A

Resultados modelos para calcular región de interés

Tabla A.1: Métricas Léxicas: Resultados del conjunto de entrenamiento por clase de la *región de interés*, usando 5-fold cross validation sobre modelo Naive Bayes. 60.19 %

Class	FP Rate	Precision	Recall	F-measure	ROC Area
0	0.607	0.556	0.829	0.666	0.662
1	0.171	0.715	0.393	0.507	0.662
W. Avg.	0.379	0.639	0.602	0.583	0.662

Tabla A.2: Métricas Léxicas: Resultados del conjunto de prueba por clase de la *región de interés*, usando 5-fold cross validation sobre modelo Naive Bayes. 60.87 %

Class	FP Rate	Precision	Recall	F-measure	ROC Area
0	0.638	0.598	0.822	0.693	0.631
1	0.178	0.637	0.362	0.462	0.631
W. Avg.	0.425	0.616	0.609	0.586	0.631

Tabla A.3: Métricas Léxicas: Resultados del conjunto de entrenamiento por clase de la *región de interés*, usando 5-fold cross validation sobre modelo Multilayer Perceptron. 65.53 %

Class	FP Rate	Precision	Recall	F-measure	ROC Area
0	0.349	0.634	0.660	0.647	0.701
1	0.340	0.676	0.651	0.663	0.701
W. Avg.	0.344	0.656	0.655	0.655	0.701

Tabla A.4: Métricas Léxicas: Resultados del conjunto de prueba por clase de la *región de interés*, usando 5-fold cross validation sobre modelo Multilayer Perceptron. 64.34 %

Class	FP Rate	Precision	Recall	F-measure	ROC Area
0	0.358	0.676	0.644	0.660	0.700
1	0.356	0.610	0.642	0.626	0.700
W. Avg.	0.357	0.645	0.643	0.644	0.700

Tabla A.5: Métricas Red: Resultados del conjunto de entrenamiento por clase de la *región de interés*, usando 5-fold cross validation sobre modelo Naive Bayes. 53.27 %

Class	FP Rate	Precision	Recall	F-measure	ROC Area
0	0.844	0.506	0.943	0.659	0.585
1	0.057	0.749	0.156	0.258	0.585
W. Avg.	0.434	0.633	0.533	0.450	0.585

Tabla A.6: Métricas Red: Resultados del conjunto de prueba por clase de la *región de interés*, usando 5-fold cross validation sobre modelo Naive Bayes. 58.30 %

Class	FP Rate	Precision	Recall	F-measure	ROC Area
0	0.820	0.568	0.932	0.706	0.638
1	0.068	0.695	0.180	0.285	0.638
W. Avg.	0.472	0.627	0.583	0.511	0.638

Tabla A.7: Métricas Red: Resultados del conjunto de entrenamiento por clase de la *región de interés*, usando 5-fold cross validation sobre modelo Multilayer Perceptron. 58.24 %

Class	FP Rate	Precision	Recall	F-measure	ROC Area
0	0.534	0.549	0.709	0.619	0.643
1	0.291	0.636	0.466	0.538	0.643
W. Avg.	0.407	0.594	0.582	0.577	0.643

Tabla A.8: Métricas Red: Resultados del conjunto de prueba por clase de la *región de interés*, usando 5-fold cross validation sobre modelo Multilayer Perceptron. 58.90 %

Class	FP Rate	Precision	Recall	F-measure	ROC Area
0	0.202	0.700	0.409	0.516	0.669
1	0.591	0.538	0.798	0.643	0.669
W. Avg.	0.383	0.625	0.589	0.575	0.669

Tabla A.9: Métricas Léxicas y Red: Resultados del conjunto de entrenamiento por clase de la *región de interés*, usando 5-fold cross validation sobre modelo Naive Bayes. 53.61 %

Class	FP Rate	Precision	Recall	F-measure	ROC Area
0	0.845	0.508	0.951	0.663	0.646
1	0.049	0.777	0.155	0.258	0.646
W. Avg.	0.430	0.648	0.536	0.452	0.646

Tabla A.10: Métricas Léxicas y Red: Resultados del conjunto de prueba por clase de la *región de interés*, usando 5-fold cross validation sobre modelo Naive Bayes. 58.55 %

Class	FP Rate	Precision	Recall	F-measure	ROC Area
0	0.821	0.569	0.937	0.708	0.639
1	0.063	0.711	0.179	0.286	0.639
W. Avg.	0.470	0.635	0.586	0.512	0.639

Tabla A.11: Métricas Léxicas y Red: Resultados del conjunto de entrenamiento por clase de la *región de interés*, usando 5-fold cross validation sobre modelo Multilayer Perceptron. 64.92 %

Class	FP Rate	Precision	Recall	F-measure	ROC Area
0	0.384	0.621	0.686	0.652	0.704
1	0.314	0.681	0.616	0.647	0.704
W. Avg.	0.348	0.652	0.649	0.649	0.704

Tabla A.12: Métricas Léxicas y Red: Resultados del conjunto de prueba por clase de la *región de interés*, usando 5-fold cross validation sobre modelo Multilayer Perceptron. 62.94 %

Class	FP Rate	Precision	Recall	F-measure	ROC Area
0	0.273	0.698	0.546	0.612	0.702
1	0.454	0.580	0.727	0.645	0.702
W. Avg.	0.357	0.643	0.629	0.628	0.702

Apéndice B

Código Python para extracción de relaciones entre usuarios mediante la API de Twitter

```
import time
import os
import sys
import json
import argparse
import key_twitter
from queries import *

enc = lambda x: x.encode('ascii', errors='ignore')

# The consumer keys can be found on your application's Details
# page located at https://dev.twitter.com/apps (under "OAuth settings")
CONSUMER_KEY = key_twitter.CONSUMER_KEY
CONSUMER_SECRET = key_twitter.CONSUMER_SECRET

# The access tokens can be found on your applications's Details
# page located at https://dev.twitter.com/apps (located
# under "Your access token")
ACCESS_TOKEN = key_twitter.ACCESS_TOKEN
ACCESS_TOKEN_SECRET = key_twitter.ACCESS_TOKEN_SECRET

# == OAuth Authentication ==
#
# This mode of authentication is the new preferred way
# of authenticating with Twitter.
auth = tweepy.OAuthHandler(CONSUMER_KEY, CONSUMER_SECRET)
auth.set_access_token(ACCESS_TOKEN, ACCESS_TOKEN_SECRET)

api = tweepy.API(auth)

def load_users(limit, offset):
    users = getUsers_Comuna(limit, offset)
    total = len(users)
    c = 1
```

```
    for user in users:
        id = user[0]
        print(str(c) + "/" + str(total) + " procesando id_user: " + str(id))
        while True:
            try:
                user = api.get_user(id)
                followers = user.followers_ids()
                total_followers_count = user.followers_count
                followers_count = len(followers)
                updateRelationUser(id, followers_count,
                    json.dumps(followers), total_followers_count)
                break
            except tweepy.RateLimitError as e:
                print(e)
                print("Wait for it")
                time.sleep(60 * 15 + 15)
            except tweepy.TweepError as e:
                if str(e) == "Not authorized.":
                    print("Can''t access user data - not authorized.")
                elif e.args[0][0]['code'] == 50:
                    print("User not found")
                else:
                    print(e)
                updateRelationUserError(id, str(e))
                break
            except StopIteration:
                break
        c += 1

if __name__ == '__main__':
    ap = argparse.ArgumentParser()
    ap.add_argument("-l", "--limit", required=True, type=int,
        help="how many users will you process?")
    ap.add_argument("-o", "--offset", required=True, type=int,
        help="since users start?")
    args = vars(ap.parse_args())

    limit = int(args['limit'])
    offset = int(args['offset'])

    print("LIMIT " + str(limit) + " OFFSET " + str(offset))

    t_ini = time.time()

    load_users2(limit, offset)

    t_fin = time.time()
    t_total = t_fin - t_ini
    print('El tiempo de ejecucion fue: ' + str(round(t_total, 2)) + ' segundos')
```