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

PREDICCIÓN DE UBICACIÓN FRECUENTE A NIVEL REGIONAL DE USUARIOS CHILENOS DE TWITTER

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERA CIVIL INDUSTRIAL

MARÍA IGNACIA CAAMAÑO LOBOS

PROFESOR GUÍA:
JUAN DOMINGO VELÁSQUEZ SILVA

MIEMBROS DE LA COMISIÓN:
ASTRID CONTRERAS FUENTES
ROCÍO BELÉN RUIZ MORENO

SANTIAGO DE CHILE
2018

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE: INGENIERA CIVIL INDUSTRIAL
POR: MARÍA IGNACIA CAAMAÑO LOBOS
FECHA: 29 DE OCTUBRE 2018
PROF. GUÍA: JUAN DOMINGO VELÁSQUEZ SILVA

PREDICCIÓN DE UBICACIÓN FRECUENTE A NIVEL REGIONAL DE USUARIOS CHILENOS DE TWITTER

El presente trabajo de título tiene como objetivo diseñar y construir modelos de clasificación para predecir cuáles usuarios de *Twitter* viven en Chile y en cuál de las quince regiones administrativas de Chile habitan; utilizando información pública del perfil y el contenido que emiten, mediante algoritmos de Machine Learning.

Las redes sociales son uno de los medios de comunicación más utilizados en el mundo hoy en día en donde segundo a segundo se reciben millones de datos. Si hablamos de *Twitter*, este contiene a diario más de 328 millones de usuarios activos en todo el mundo que publican cerca de 6.000 tweets por segundo. Todos estos datos que se comparten son de gran utilidad para poder caracterizar de mejor forma a los usuarios. Una de estas características es la geolocalización, la cual está siendo cada vez más utilizada para conocer mejor a los clientes y usuarios. Pero, obtener este atributo para cada usuario no es tarea fácil ya que este dato, por lo general, no es público o es incierto.

En particular, el proyecto SONAMA y proyecto OpinionZoom del Web Intelligence Center, buscan geolocalizar a los usuarios chilenos de *Twitter* a nivel regional para utilizar esta característica dentro de sus investigaciones y poder mejorar los niveles de granularidad que están alcanzando con respecto a la geolocalización de las personas.

Para esto, a partir de la API REST de *Twitter* y de una encuesta realizada en el WIC, se extraen datos para construir bases de entrenamiento etiquetadas para diseñar y entrenar dos modelos de clasificación con el fin de que uno identifique a los usuarios chilenos de *Twitter* y el segundo identifique la región en que cada usuario chileno vive. Para ambos modelos se analiza el desempeño de tres algoritmos diferentes.

El modelo país, mediante Stochastic Gradient Descent, logra un AUC de 99,89 % y un *F1-Score* de la clase positiva de 98,95 % tras validación cruzada de 5-iteraciones, el cual supera los resultados de la heurística de clasificación que actualmente se utiliza. Por otro lado, el modelo región, mediante Stochastic Gradient Descent, logra un *F1-Macro* de 42,54 % y *Accuracy* de 40,73 % tras validación cruzada de 5-iteraciones. Resultado el cual, si bien bajo, mejora la situación actual con respecto al 6,67 % que entrega la aleatoriedad.

Finalmente, se valida la hipótesis de investigación en su totalidad ya que es posible determinar los usuarios chilenos de *Twitter* y su región frecuente utilizando exclusivamente el contenido que se publica y atributos de contexto del usuario. Tanto el modelo país como el modelo región superan la situación actual, por lo que sus resultados quedan como *baseline* para próximas investigaciones.

'Un día sin reír es un día perdido'
Charles Chaplin

Agradecimientos

Por fin llegó el momento de terminar la carrera y me gustaría agradecer a todos los que de alguna u otra forma fueron parte de mi vida en este proceso.

Quiero partir agradeciéndole a mis papás, Marianella y Ricardo, por todo lo que me han entregado, por enseñarme desde (más) pequeña la importancia del esfuerzo, que las cosas no llegan solas y por estar siempre muy preocupados por mí y entregándome su cariño.

A mis padrinos, Andreita y Yeye, que han estado en todas conmigo, igual que mis papás. Apoyándome, cuestionándome, felicitándome, retándome y por sobre todo, queriéndome como una hija más. A mis primillos, Ale y Juan, por ser mis hermanos. Y también a mi Guely, la más linda e incondicional.

Al Ale, que se ha convertido en una persona muy importante en mi vida. Gracias por todo tu apoyo, por los lindos momentos, por intentar sacarme una sonrisa cuando no he tenido un buen día y por ser mi fan número 1 y mi peor crítico a la vez.

A mis amigos y amigas del colegio y de scout, que aunque algunas veces nuestros caminos se vayan separando, siguen estando ahí para un consejo, pasarlo bien o simplemente hablar. En especial al Mati, Nacha, Gilda, Vale, Dani, Olo y Pasita.

A la Feria Empresarial por marcar mi paso por la Universidad estos últimos 4 años y por hacer que conociera a grandes amigos! Gracias al 'Patio' por todos los tremendos momentos que hemos compartido, por los que se vienen y que la amistad siga por mucho tiempo más.

A mis partners de clases y de los antojos de cosas dulces a la hora que fuera, la Kari, Vania y Naty. Gracias por ser tan preocupadas siempre, por darme ánimo cada vez que lo necesito y por soportarme en esos innumerables días de enojo y estrés.

Finalmente, agradecer al WIC por permitirme realizar la memoria con ustedes y por su buena disposición siempre. Y al Pangui, mi tutor, y a mis profesores, Juan, Astrid y Rocío, por su ayuda en este proceso. Gracias también a la sección 3, a 'los wagyus', taller 1, a 'los memoriones', a pole, a scout, a mi familia y amigos en general y a todos los que tuve la posibilidad de conocer y compartir en estos casi 7 años, por los buenos momentos.

¡Gracias infinitas, por fin seré Injeniera!

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

Hoy en día, las redes sociales son uno de los medios de comunicación más utilizados en el mundo. Si bien no en todos los lugares hay acceso a una red que provea Internet o bien a un dispositivo para conectarse, esto se ha vuelto cada vez más masivo. Los usuarios de los medios y redes sociales comparten día a día una gran cantidad de información ya que las utilizas como medio de difusión, distribución e incluso de expresión de opiniones y sentimientos.

Las tecnologías de hoy nos permiten capturar datos provenientes de las redes sociales, analizarlos y obtener conclusiones interesantes sobre diversas temáticas, como por ejemplo, el análisis de sentimientos de los comentarios que los usuarios realizan, la segmentación de usuarios según sus gustos y preferencias para aplicarlo a marketing focalizado, entre otras. Cabe destacar que dichos datos se pueden obtener sí y sólo sí los usuarios de las diferentes redes sociales autorizan a cada una de éstas a que uno o más datos que ingresan sean públicos.

La geolocalización es un atributo bastante valorado en diferentes estudios sobre detección de eventos [29], recomendación de productos y/o servicios basados en la ubicación [32] [9] [27] o incluso para determinar las temáticas principales que se están tratando por zona geográfica [24]. Este atributo permite identificar dónde se ubican o ubicaron geográficamente los usuarios y así comprender sus comportamientos o los sucesos que están ocurriendo a su alrededor según dónde están ubicados. Es por esto que obteniendo la geolocalización podrían determinarse los lugares que frecuenta un usuario o podría utilizarse para determinar ubicación en tiempo real. Esta, además, tiene diferentes niveles ya que puede ser ubicación a nivel de continentes, país, estados, regiones, comunas o POI (*Point of Interest*), en que se obtiene un punto de coordenadas específico. Para cada tipo de geolocalización que se busque obtener, se requiere diferente información ya que, por ejemplo, los lugares que se frecuentan requiere historia de comportamiento del usuario en cuanto a los lugares en que ha estado y la ubicación en tiempo real sólo depende de lo que está pasando en el momento.

Para obtener la geolocalización de una persona desde un medio social[1], se pueden observar los campos en que los usuarios indican el lugar en donde viven (país, región, estado, comuna, etc), las etiquetas de ubicación que agregan los usuarios al contenido que publican (*geotag*) o también si comparten su ubicación en tiempo real. El problema radica en que gran parte de

[1]Sólo si la persona tiene su información pública

los usuarios que utilizan redes sociales no comparten públicamente este tipo de información, y si la comparten muchas veces ingresan información no verídica [18].

Sin embargo, *Twitter* se ha instaurado como una red social de tipo micro-blog en la que una gran cantidad de usuarios tiene su información pública y, por ende, la red mediante la cual se puede obtener una mayor cantidad de datos para realizar diferentes estudios, incluidos los de geolocalización. En particular, en Chile, *Twitter* no se queda atrás y en diversos estudios previos, como [6] [10] [22] entre otros, se ha constatado la riqueza de la utilización de los datos públicos que se comparten día a día en esta red.

1.1. Web Intelligence Centre

Es un centro de investigación en tecnologías Web del departamento de Ingeniería Industrial de la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile.

Misión: "¡Creando soluciones ingenieriles inteligentes! "Creamos tecnologías de información usando data science para apoyar la toma de decisiones en organizaciones que se interesan en innovar. Creemos que esta disciplina puede generar un gran impacto en la sociedad y eso nos apasiona."

Visión: Ïmpactar el mundo "Queremos ser un actor relevante en el área del data science y sus aplicaciones en la salud. Para ello contaremos al 2022 con al menos 3 proyectos transferidos en distintos centros de salud."

Objetivos:

- Publicar en las principales revistas, conferencias y editoriales relacionadas con Web Intelligence.
- Proveer a servicio profesional, excelente y rápido para todos nuestros clientes.
- Dictar cursos de orientación práctica acerca de las tecnologías de información y su aplicación en los negocios.

La presente investigación se realiza en el marco de dos proyectos que está llevando a cabo el Web Intelligence Centre (WIC). Estos son SONAMA y OpinionZoom.

Proyecto SONAMA
Plataforma que monitorea en línea la opinión y consumo sobre marihuana, tabaco y alcohol de los usuarios chilenos de Twitter.

Proyecto OpinionZoom
Plataforma enfocada en el análisis en tiempo real de las opiniones que los usuarios chilenos de Twitter están aportando en la Web sobre un tema/producto/empresa en particular que puede permitir, a las organizaciones que lo utilicen, tomar acciones inmediatas.

1.2. Descripción y justificación

Los usuarios de *Twitter* entregan una alta cantidad de información en lo que publican y comparten públicamente, sin saberlo conscientemente. Estudiar estos datos tiene un gran potencial, sobre todo por los diferentes tipos de información que se puede extraer de ellos, ya que día a día más de 1.7 millones de personas *tweetean* en Chile según un estudio de Universia [4].

Diferentes casos de estudio buscan capturar la localización geográfica de las personas para poder tomar decisiones, enfocar campañas, determinar indicadores, obtener estadísticas, etc. En el contexto de las redes sociales, como *Facebook*, *Twitter*, *Linkedin*, los usuarios tienen la posibilidad de geolocalizar sus publicaciones y a partir de dicha información, los investigadores pueden sacar ciertas conclusiones. El problema es que una gran minoría accede a utilizar públicamente la geolocalización en sus perfiles y/o publicaciones, por lo que las conclusiones que se pueden sacar a partir de la poca información georeferenciada termina siendo sesgada y por ende no es muy correcto extrapolarla.

Es por esto que en cuanto a la cantidad exacta de usuarios chilenos de *Twitter*, no es posible calcularla exactamente ya que sólo existe un campo en el perfil del usuario, llamado campo *location*, en el cual cada usuario escribe manualmente dónde vive, y no necesariamente escribirá el país/región/estado/comuna/calle correspondiente. Para poder estimar esa cantidad se han realizado una serie de estudios que buscan calcularlo, pero todos estos utilizan diferentes formas de determinarla y por lo mismo la cifras difieren unas de otras, como en [4] [1]. Dichos estudios sólo utilizan la información que se obtiene del perfil del usuario con respecto a la ubicación que esa misma persona ingresa, por lo que si alguien que vive en Chile no ingresa en el campo *location* que actualmente vive en Chile, entonces no se considerará en la base de datos.

El problema de utilizar el campo *location* para determinar la ubicación de los usuarios radica en que, de las pocas personas que ingresan este campo, sólo el 40 % ingresa algo certero [5]. Además, en cuanto a la información georeferenciada que proveen los *tweets*, menos del 1 % de éstos poseen el *geotag* que es la etiqueta de la localización en donde el *tweet* fue emitido [22].

La oportunidad de poder determinar los usuarios chilenos y la región en la que viven las personas con cuenta de *Twitter* proviene de una necesidad por caracterizar a los usuarios de manera un poco más granular en los dos proyectos antes mencionados que se están realizando en el centro.

Por un lado, está el proyecto SONAMA el cual, de acuerdo a investigaciones previas realizadas en el WIC [7] [12], comprueba que es posible identificar el consumo de marihuana y alcohol de la población en Chile. Esto se realizó a partir del contenido que los usuarios chilenos de *Twitter* generan en dicha red social. Ambas investigaciones demuestran lo anterior logrando replicar valores similares de algunas métricas de la Encuesta Nacional de Drogas y Alcohol que realiza el Servicio Nacional para la Prevención y Rehabilitación del Consumo de Drogas y Alcohol (SENDA) cada dos años, y que determinan el comportamiento a nivel nacional con respecto a la marihuana y alcohol. El tema es que dichos resultados sólo fueron

posibles de validar a nivel país y no a nivel regional, que es la forma en que la Encuesta Nacional de Drogas y Alcohol entrega los datos, ya que no se tiene la información sobre las regiones en que viven los usuarios chilenos. Entonces, al investigar un modelo que permita identificar la región en la que viven los usuarios chilenos de *Twitter* permitiría que el proyecto SONAMA pueda validar o rechazar la hipótesis de que las métricas a nivel regional de la Encuesta Nacional de Drogas y Alcohol también son replicables a partir de los datos públicos que generan los usuarios chilenos de *Twitter*. En caso de validarse dicha hipótesis, SONAMA podría convertirse en una herramienta de utilidad para el SENDA para ir monitoreando el comportamiento con respecto a drogas y alcohol de los chilenos, por región. Incluso, podría utilizarse para medir el impacto en el corto plazo que tenga alguna medida y/o política pública en alguna región en particular o a nivel país, relacionada al consumo de drogas y alcohol.

Por otro lado, está el proyecto OpinionZoom que es una plataforma que actualmente ofrece a sus clientes poder conocer en tiempo real lo que se está diciendo sobre su marca o sobre algún tema de interés. Esto les permite a los clientes poder gestionar de mejor manera sus productos, campañas publicitarias, difusión focalizada, entre otros. Este análisis de opiniones indica:

- ¿Quiénes están hablando sobre la marca/producto/tema?
- Polaridad de las opiniones (positivas, negativas o neutras).
- ¿Quiénes lideran las opiniones?
- Tópicos de interés de los que opinan.
- Identificación de reclamos sobre marca/producto/tema.
- Alertas sobre cambios de opinión.

Hoy en día, toda esta información se basa en datos agregados en cuanto a ubicación geográfica, es decir, no es posible determinar si este análisis de opiniones tiene tendencia en algún lugar en particular o si las opiniones positivas se concentran más en un lugar que en otro. Es por esto que poder determinar la geolocalización regional de los usuarios podría aportar en el análisis de opiniones que actualmente ofrece OpinionZoom ya que se podría tener información desagregada por región. Esto permitiría agregar valor al servicio que actualmente se ofrece a los clientes de OpinionZoom ya que la plataforma no sólo indicaría lo que se está diciendo a nivel país, si no que indicaría lo que se está opinando con respecto a cierta empresa/tema a nivel de cada región. Por consecuencia, cada región podría tener a modo general opiniones con diferentes polaridades, por lo que los clientes de OpinionZoom podrían tomar acciones más concretas y eficientes en cada región.

Entonces, para poder caracterizar a los usuarios chilenos de *Twitter* en base a su ubicación más frecuente para así determinar su región de residencia, es fundamental poder detectar cuáles son las cuentas de *Twitter* que pertenecen a usuarios que viven en Chile y en segundo lugar detectar en qué región de Chile viven dichas personas.

En cuanto a la detección de usuarios que viven en Chile, actualmente para ambos proyectos se utiliza una heurística que recorre usuarios de *Twitter* y mediante la identificación de la palabra 'Chile' u otros nombres de regiones, cuidades o comunas del país en el campo location de *Twitter*, captura a los usuarios chilenos. Pero, se busca evaluar la opción de utilizar también

otros atributos para la extracción de usuarios de *Twitter* que con alta probabilidad vivan en Chile, para así contar con una base más amplia y precisa de chilenos. Esto ya que se debe considerar que, tal como se mencionó anteriormente, en el Universo de *Twitter* sólo el 40 % de los que ingresan información a su campo *location* tienen información verídica.

1.3. Hipótesis de investigación

Al ser *Twitter* una red social con alto impacto en Chile y poseer una gran cantidad de información pública de los usuarios, se tiene como hipótesis que es posible determinar la ubicación frecuente a nivel regional de los usuarios de *Twitter* que viven en Chile, identificando primero cuáles son los usuarios que actualmente viven en Chile para luego identificar la región, utilizando la información pública del perfil y el contenido de lo que publica cada usuario en *Twitter*, mediante el uso de algoritmos de *Machine Learning*.

1.4. Objetivos

1.4.1. Objetivo General

Diseño y construcción de modelos de clasificación para predecir los usuarios de Twitter que viven en Chile y predecir en cuál de las quince regiones administrativas de Chile habitan, a partir de la información pública de su perfil y del contenido que publican, mediante algoritmos de Machine Learning.

1.4.2. Objetivos Específicos

- Construir set de datos de entrenamiento y testeo para modelo de clasificación que identifica usuarios chilenos en *Twitter*.
- Construir set de datos de entrenamiento y testeo para modelo de clasificación que identifica región en que habitan los usuarios chilenos de *Twitter*.
- Entrenar, evaluar desempeño y evaluar mejora con respecto al método anterior de modelo de clasificación para obtener cuáles cuentas de *Twitter* pertenecen a usuarios que viven en Chile.
- Entrenar y evaluar desempeño de modelo de clasificación que permita determinar ubicación frecuente regional de los usuarios chilenos.
- Validar los resultados obtenidos de ambos modelos para determinar la utilidad de los modelos construidos.

1.5. Metodología

Para llevar a cabo los objetivos planteados es necesario tratar el problema con dos focos, la clasificación de los usuarios de *Twitter* que viven en Chile y la determinación de la región de residencia de los usuarios chilenos. Ambos focos se trabajarán mediante dos fases metodológicas:

- **Fase 1 - Investigación:** Se realiza proceso investigativo sobre el estado del arte de la inferencia de geolocalización en *Twitter* mediante *text mining* y *data mining*, para así poder identificar el tipo de información que se ha utilizado junto con las diferentes metodologías que se han implementado. Esto con el fin de generar una propuesta de solución complementando diferentes investigaciones que han llevado a cabo este tipo de problema a partir de la detección de aquellas técnicas que hayan tenido mejor desempeño y aplicando principales descubrimientos que se hayan estudiado.

  En esta fase se define que el enfoque de los modelos de clasificación a entrenar es relacionado al contenido y contexto de los usuarios en *Twitter*, y que los algoritmos a utilizar son Naïve Bayes, Random Forest y Support Vector Machine utilizando Stochastic Gradient Descent.

  Se determina el enfoque de la investigación en relación al contenido y contexto de los usuarios ya que de las investigaciones estudiadas en 2.9 eran en relación al contenido, contexto, red de seguidores y seguidos, contenido y contexto o una mezcla de todas. Pero, generalmente la investigación como punto de partida en cada país era en relación al contenido y/o contexto, por lo que se decide comenzar con ese enfoque de manera de obtener un *baseline* para luego continuar mejorando la capacidad de geolocalización en Chile en base a la presente investigación.

- **Fase 2 - Metodología KDD:** Se trabajará en base a la metodología KDD siguiendo dicho proceso para cada uno de los modelos a entrenar. Los datos para entrenamiento, validación y testeo de modelos se obtienen de 2 fuentes: la API de *Twitter* y los resultados de una encuesta realizada por el WIC a usuarios chilenos de *Twitter* sobre drogas y alcohol.

  Para construir los set de datos de entrenamiento se identifican cuentas de usuarios de *Twitter* que viven en Chile, cuentas de usuarios de *Twitter* que viven en otros países del mundo y cuentas de usuarios que viven en cada una de las regiones de Chile. Para determinar a los usuarios del mundo se identifican primero aquellos usuarios que tienen geolocalización activada y según los países en que fueron detectados su *tweets*, el usuario se asocia al país en el que tuvo mayor frecuencia de aparición. En cuanto a la determinación de los chilenos, además de la forma anterior se identifican mediante el campo *location* si es que contiene la palabra 'Chile'; y también aquellos que respondieron que viven actualmente en Chile en la encuesta. Finalmente, para la determinación de la región de los usuarios chilenos, se hace mediante el campo *location* identificando el nombre de las regiones, provincias y principales ciudades de cada región; y también se hace mediante los lugares a los que están asociados los *tweets* geolocalizados.

  Luego de obtener los usuarios necesarios, se extraen sus *tweets* y *retweets* de un año.

De esta forma se logran obtener tanto las variables de contexto como la variable de contenido de cada usuario.

Con los datos ya seleccionados, se preprocesan para limpiarlos y obtener sólo aquellas variables que se utilizarán para la construcción del modelo, las cuales son zona horaria, lenguaje de la interfaz, campo *location* y conjunto de *tweets* y *retweets* emitidos en un año.

Luego, se procede a realizar *text mining* y *data mining* para elegir el algoritmo que mejor modele el problema. Finalmente se construyen las métricas de desempeño para evaluar qué tan certeros son los modelos, junto con establecer una métrica de comparación del modelo de clasificación de usuarios chilenos versus el algoritmo que actualmente se utiliza para identificar a los usuarios de Chile, de manera que se pueda saber cuantitativamente si el modelo propuesto es mejor que la situación actual. Además, para validar cuantitativamente si se cumple la hipótesis investigativa de poder determinar la región en que vive un usuario chileno mediante algoritmos de *machine learning*

1.6. Resultados esperados

Luego de la realización del proyecto se busca tener como resultado la validación de un modelo de clasificación que permita determinar si un usuario de *Twitter* vive en Chile o no. Además de un segundo modelo de clasificación que permita determinar la región de Chile en la que viven los usuarios de *Twitter* que fueron identificados como habitantes de Chile.

1.7. Estructura de la memoria

El capítulo 2 corresponde al Marco teórico y conceptual, el cual contiene todo lo relacionado al estado del arte de la inferencia de la geolocalización, para poder definir más adelante la metodología y técnicas a utilizar en el proyecto. Además, contiene la descripción e información relevante sobre los conceptos claves para entender el contexto y desarrollo del trabajo.

Más adelante, en el capítulo 3 se detalla el proceso de la construcción de los set de entrenamiento para los dos modelos a realizar. Se describe cómo se obtienen los datos, las fuentes y la elección de los usuarios para formar parte de la base, junto con el proceso de etiquetado para definir los usuarios que son chilenos o extranjeros, y si son chilenos, la región en la que viven.

A continuación, en el capítulo 4 se presenta el modelamiento de los dos problemas que se abordan explicitando resultados para los tres algoritmos que se prueban y utilizando diferentes atributos para entrenar cada modelo, logrando así identificar diferentes variaciones de los modelos para determinar cuál es la que reporta un mejor desempeño tanto a nivel de modelo país como a nivel de modelo región.

Luego, en el capítulo 5 se analizan y discuten los resultados obtenidos en la fase de modelamiento, eligiendo el modelo que más se ajusta a cada problema.

Finalmente, se concluye en base a los resultados obtenidos y los análisis expuestos, para poder determinar si se logra validar la hipótesis de investigación.

Capítulo 2

Marco teórico y conceptual

En el presente capítulo se dan a conocer conceptos relevantes que permiten dar contexto al trabajo realizado en los capítulos posteriores de manera que se pueda obtener una completa comprensión de lo desarrollado.

2.1. World Wide Web

Es un sistema interconectado de páginas web públicas a las cuales se puede acceder a través de internet, que surge como idea en 1989 por la necesidad de acceder a cierta información desde distintos lugares sin tener que hacerlo a través del mismo computador. Tim Berners-Lee, un profesional de las ciencias de la computación británico, propuso el poder compartir la información explotando la tecnología de hipertexto a través de internet, medio el cual se estaba comenzando a desarrollar.

Es así como nace la WWW, la cuál está compuesta por: [2] [26]

- **Protocolo HTTP (*HyperText Transfer Protocol*):** Permite recuperar la información de los recursos que están linkeados de toda la web, dirigiendo la transferencia de datos que se realiza entre el servidor y cliente.
- **URI (Uniform Resource Identifier) o URL:** es una dirección única que se utiliza para identificar cada recurso que está en la web y poder acceder a el.
- **HTML (HyperText Markup Language):** es el lenguaje que le da formato a la web para poder publicarlos.

A lo largo del tiempo, la web se ha ido desarrollando generando así diferentes versiones.

La Web 1.0
Consiste en una web de sólo lectura en el que el cliente sólo recibe el contenido de la página de manera estática por lo que no puede interactuar con ella. La única forma de poder interactuar era a través de correo electrónico, *chat* o foros. Todo estaba bajo el control de un *webmaster*.

La Web 2.0
Este termino lo determina Tom O'Reilly en el año 2004 para marcar el cambio entre la página estática y una basada en usuarios activos, dando paso a los *blogs*, redes sociales, páginas web creadas por usuarios, entre otras. Dado esto, quienes generan contenido ya no son sólo los *webmaster* si no que también los mismos usuarios.

2.2. Medios sociales

Son plataformas de comunicación en las que el contenido es creado y compartido por usuarios en línea. Éstas son parte fundamental de la web 2.0 ya que mediante el uso de su tecnología los medios sociales lograron que los usuarios pudiesen interactuar entre ellos, generando comunidades o grupos en que se pudiesen vincular en un mismo espacio virtual. Llegaron a transformar las formas de comunicación haciendo que los usuarios compartan información de diferentes formas: publicaciones escritas, fotos, imágenes temporales, videos, geolocalización, música, entre otros. Algunas de los tipos de medios sociales que existen actualmente son los blogs, redes sociales, microblogs, entre otros. Según la RAE[1], una red social es una plataforma digital de comunicación global que pone en contacto a un gran número de usuarios; un blog es un sitio web que incluye, a modo de diario personal de su autor o autores, contenidos de su interés, actualizados con frecuencia y a menudo comentado por los lectores. Por otro lado, el microblog es un blog en donde los usuarios publican mensajes breves que pueden ir acompañados de enlaces, imágenes o archivos audiovisuales.

Existen redes sociales de temas generales como *Facebook*, de transporte como *Waze*, de turismo como *Tripadvisor*, de contenido audiovisual como *Instagram*, de mensajería como *Whatsapp*, de música como *Spotify*, de citas como *Tinder*, de contactos laborales como *Linkedin*, entre otras. En el caso de *Twitter*, este medio se considera como un *microblogging* pero también como una red social. *Microblog* ya que las publicaciones que se realizan tienen un tamaño máximo, y red social debido a que los usuarios pueden relacionarse bidireccionalmente con otros usuarios mediante relaciones asimétricas dado que un usuario puede seguir a otro pero no necesariamente se siguen entre sí.

2.2.1. Twitter

Red social y *microblogging* que se utiliza como medio de comunicación en la que cada persona puede crearse una cuenta y publicar mensajes ilimitados de 280 caracteres máximo cada uno. Dichas publicaciones se llaman *tweets* y aquellas publicaciones en las que un usuario comparte el *tweet* que realizó otro usuario se llama *retweet*. Cada usuario tiene seguidores, que son las personas que desean ver lo que dicho usuario publica, mientras que los seguidos son las personas que el propio usuario desea visualizar lo que publican. Los mensajes que cada usuario publica, ya sea un *tweet* o un *retweet*, llegan a la página principal de *Twitter* de todos sus seguidores.

[1] http://www.rae.es

Cada *tweet* puede contener diferentes características de *Twitter*:

Hashtag: Son las palabras que se comparten en un texto que van antecedidas por un signo "#". Estos elementos se utilizan para marcar y enfatizar el tema principal del *tweet*, y el usuario que publica es libre de poner el hashtag que desee. Mediante esta característica se pueden buscar *tweets* sobre la misma temática al buscar un hashtag en particular. Ejemplo: El hashtag #Cencosud debería contener publicaciones que se han realizado sobre la empresa Cencosud.

Menciones: Esta característica se utiliza para mencionar a otros usuarios de *Twitter* de manera que se le avise a dicho usuario que está siendo mencionado. Se reconocen por utilizar el signo "@"previo al nombre de usuario. Cuando se *retweetea* una publicación se agrega automáticamente una mención del usuario que publicó el *tweet* que está siendo *retweeteado*.

Esta red social se utiliza principalmente para publicar opiniones, comentarios, reclamos, pensamientos, etc. A diferencia de *Facebook*, por ejemplo, que tiene como uso principal las publicaciones de las actividades del día a día.

Tipo de información de *Twitter*

Twitter cuenta con diferentes tipos de información. En primer lugar, está la información personal que cada usuario comparte en su perfil, y esta se puede tener pública o privada. También está la información sobre los mensajes (*tweets*) que los usuarios publican en la plataforma y la información sobre las personas que siguen a cada usuario y las que sigue un usuario en particular. La localización geográfica a partir de los *tweets* se ha estudiado de 3 formas: la ubicación que se menciona en el *tweet*, la ubicación en la que fue emitida el *tweet* y la/s ubicación/es que el usuario frecuenta (típicamente el hogar). Esto puede verse desde el foco tanto del contenido del *tweet* como de las redes de usuarios de *Twitter*. [33]

2.3. Application Programming Interfaces (API)

Tal como lo indica su nombre en inglés, es una interfaz de programación de aplicaciones que permite que un software y una persona interactúen y se comuniquen para realizar intercambio de datos. Es un conjunto de funciones, procedimientos o incluso métodos que ofrece una biblioteca para ser utilizado como capa de abstracción por otro software. Existen API's de servicios web, API's basadas en bibliotecas, API's basadas en clases y API's de funciones en sistemas operativos. Dentro de las primeras, el intercambio de información se realiza entre una web y una aplicación. Un ejemplo de esta es la API de Twitter.

2.3.1. APIs de *Twitter*

Twitter cuenta con 3 APIs públicas que son API REST, API Streaming y Search API.Para poder acceder al *core* de los datos de *Twitter* [2], como los *tweets*, perfiles de usuarios, seguidores, etc, se utiliza la API REST que entrega los datos en formato JSON. También se permite crear nuevos *tweets* mediante la API REST. Por otro lado, la API Streaming se puede utilizar para acceder a algunos *tweets* públicos que se están emitiendo en tiempo real y mediante la Search API se pueden obtener *tweets* de hasta 7 días atrás que cumplan con la consulta que se le realice.

Estructura de datos de *Twitter*

Los datos a los que se puede acceder mediante las API's se obtienen en formato JSON los cuales se basan en pares de llave y su valor. En el JSON de *Twitter* se entrega toda la información sobre el *tweet*, y dentro de ésta información se pueden identificar 4 objetos principales: *User* entrega toda la metadata sobre la cuenta del usuario, *Places* entrega metadata sobre la geolocalización del *tweet* ; *Entities* entrega los vectores de datos relacionados a los *hashtags*, menciones que se le realice a otro usuario en el *tweet*, símbolos, archivos multimedia o URLs, y *Extended entities* entrega información sobre los archivos que se adjunten al *tweet*, tal como lo hace *entities* solo que en esta se especifica si el contenido adjunto corresponde a una foto, video o GIF y además caracteriza a los 4 archivos que se pueden adjuntar en el *tweet* mientras que *entities* sólo entrega información sobre 1.

Geolocalización en los *tweets*

Los *tweets* que publican los usuarios pueden estar asociados a una ubicación mediante un punto específico que indica coordenadas de la longitud y latitud o mediante un lugar que pertenece a un área determinada por un punto de coordenadas sur-oeste y otro punto de coordenadas nor-este llamada *bounding box*, a la cual pertenece el lugar que se está indicando. Éstos se ven reflejados en el objeto *'coordinates'* del JSON y en el objeto *'places'*, respectivamente. Siempre que un *tweet* contenga *geo-tag* aparecerá *'places'* mientras que *'coordinates'* sólo aparecerá cuando el *tweet* esté asociado a una ubicación exacta.

- ***Places***
  Si un usuario tiene activada la opción geo_enabled, entonces tendrá la posibilidad de asociar sus *tweets* a algún lugar cada vez que los publique. La información de este lugar aparece en el JSON correspondiente a cada *tweet* y en el se pueden identificar las variables indicadas en la figura 2.1.
- ***Coordinates***
  Indica una ubicación exacta mediante un punto de coordenadas y éstas se pueden visualizar en el JSON, en el caso de que el *tweet* esté asociado a una ubicación exacta, en las siguientes variables del objeto *'coordinates'*:

[2]Con un máximo de 3200 *tweets* por usuario

Variable	Tipo	Descripción
id	String	Id que representa el lugar
url	String	URL donde se encuentra metadata adicional del lugar
place_type	String	Tipo de ubicación que representa el lugar. Ej: ciudad
name	String	Nombre corto del lugar
full_name	String	Nombre largo del lugar
country_code	String	Código corto del país en que se encuentra el lugar
country	String	Nombre del país en que se encuentra el lugar
bounding box	Object	Área delimitadora del lugar con sus respectivas coordenadas y el tipo de área que se esta representando
attributes	Object	Nulo al utilizar PowerTrack, APIs de búsqueda de 30 días y de archivo completo, y Volume Streams

Figura 2.1: Variables del objeto *'places'*

Variable	Tipo	Descripción
coordinates	Collection of float	Longitud y latitud de la ubicación del tweet
type	String	Tipo de la data codificada en coordinates

Figura 2.2: Variables del objeto *'coordinates'*

Geolocalización según el perfil

En el perfil de usuario, cada persona puede ingresar manualmente su ubicación. Este campo es el campo *location* el cual no necesariamente todos los usuarios ingresan y el formato en el que se indica la ubicación es libre, por lo que un usuario podría indicar otro lugar de residencia o escribir algo que no pueda ser georeferenciado.

2.4. Extracción de datos

2.4.1. *Crawler*

Es un programa que se utiliza en la World Wide Web para recorrer página por página buscando lo que el usuario indique de una forma automatizada. Esto se implementa mediante un tipo de algoritmo llamado búsqueda *breadth-first* que recorre y busca elementos en grafos y realiza esta búsqueda sin información. Su función principal es poder recolectar de una manera automática las páginas web que sean de interés, y este recorrido a grandes rasgos lo realiza a partir de páginas web iniciales, llamadas semillas, que son las primeras en ser analizadas, y deja en cola todos los hipervínculos que se identifiquen en dichas páginas. Luego, analiza las páginas en cola y vuelve a agregar a la cola los hipervínculos que detecte, y así sucesivamente.

En el caso de Twitter, esta red social al igual que la World Wide Web se puede modelar como un grafo considerando que cada nodo es un usuario y las aristas van dirigidas de un nodo a otro si es que el usuario del nodo inicial sigue al usuario del nodo final. En base esto, se pueden recorrer los usuarios de *Twitter* mediante un algoritmo *breadth-first* de una manera

automática utilizando ciertos usuarios como semillas.

2.5. *Ciencia de datos*

Se refiere al estudio de datos relacionado al manejo, integración, arquitecturas, aprendizaje automático, visualización, entre otras; para así poder tomar decisiones de negocios basadas en datos.

2.5.1. Minería de datos

Es una técnica utilizada para buscar patrones no visibles en una gran cantidad de datos que pueden ser de gran utilidad para predecir ciertas cosas. En otras palabras, permite encontrar información y conocimiento relevante acerca de datos que por si solos no dicen mucho.

Metodología KDD

Es un proceso interactivo e iterativo que utiliza los métodos de Data Mining para extraer conocimiento de los datos. Considera 5 etapas que son: [25]

- **Selección:** Se obtiene una muestra del set de datos con que se trabajará, acotándolo sólo a las variables y registros que se requerirán.
- **Pre-procesamiento:** Limpieza de la muestra de datos con la que se trabajará.
- **Transformación:** Etapa en que se transforman los datos para crear nuevas variables, reducir dimensionalidad, adaptarlos para utilizarlos en los modelos, etc.
- **Data Mining:** Consiste en la búsqueda de patrones mediante un modelo particular que se ajuste a lo que se necesita dependiendo del objetivo y de los datos que se tienen.
- **Interpretación/Evaluación:** interpretación de los patrones/correlaciones obtenidos junto con la evaluación del modelo en base a métricas de desempeño.

2.5.2. Minería de texto

Es una de las técnicas que se utilizan dentro de la minería de datos. Es un tema particular dado que ésta se enfoca en extraer conocimiento e información a partir de documentos de texto. Una de las grandes diferencias con las técnicas clásicas de la minería de datos, es que ésta siempre posee información no estructurada. Para trabajar con esta técnica se puede utilizar la metodología KDD, pero lo que se realiza en cada etapa difiere bastante de lo que se realiza típicamente en datos de una base de datos dado que se suelen utilizar diccionarios con palabras claves para trabajar los documentos.

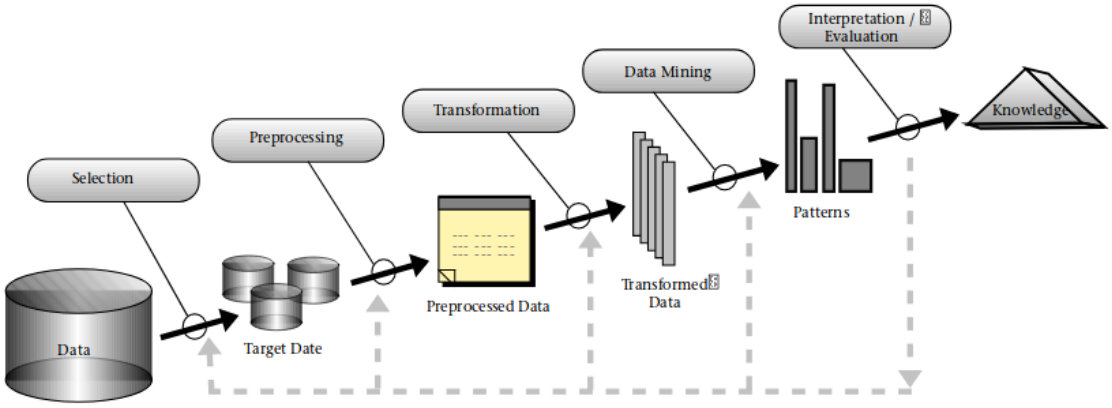

Figura 2.3: Pasos del proceso KDD
Fuente: [15]

2.5.3. *Machine learning*

El aprendizaje de máquinas o aprendizaje automático lo que realiza principalmente es el desarrollo de técnicas que permitan aprender de los datos para adaptar ese conocimiento en futuras aplicaciones, de manera que no solo se pueda explicar lo que pasó anteriormente si no que también pueda aplicarse para anticiparse a lo que probablemente ocurrirá.

Si bien se hace una cierta diferencia entre minería de datos y aprendizaje automático, finalmente ambas convergen a lo mismo en relación al descubrimiento de patrones que permitan aplicarse en nuevo set de datos para apoyar la predicción. Ya sea en problemas de segmentación, de motores de búsqueda, motores de recomendación, clasificación, etc.

Aprendizaje Supervisado

Los sistemas basados en el aprendizaje supervisado consisten en que la base de datos con la cual se entrenan los modelos cuenta con los valores/nombres de la variable a predecir, por lo tanto al momento de medir la eficiencia del modelo entrenado se puede comparar el resultado predicho con el resultado real.

Para comparar el resultado predicho con el resultado real, es necesario contar con un set de datos diferente a los que se utilizaron para el entrenamiento ya que para medir la eficiencia real del modelo se necesita que a partir de datos que el modelo no conoce previamente, este pueda predecir correctamente. A este set de datos se le llama set de testeo, que es en el cual se prueba y se evalúe el desempeño del modelo entrenado.

Aprendizaje No Supervisado

Este método consiste en que los datos con los que se entrena el modelo no cuentan con una variable que indique el resultado al que se debe llegar, por lo que todas las variables de entrada son variables aleatorias para el modelo. Este se utiliza principalmente para obtener códigos factoriales y luego utilizarlos en técnicas con aprendizaje supervisado y también para realizar clusterización.

K-Fold Cross Validation

Este método, llamado validación cruzada, se utiliza en el proceso de entrenamiento de datos que permite asegurarse de que no exista sobreajuste al entrenar el modelo con el set dado. Esto se logra ya que validación cruzada particiona un set de datos en k partes iguales, luego en la primera iteración entrena un modelo con (k-1) particiones y testea en el restante, luego en la siguiente iteración deja otra partición para testeo y entrena en las particiones restantes, y así sucesivamente hasta llegar a las k-iteraciones.

Mediante este método se valida la independencia de la partición que se utilice para entrenar a partir de las métricas que este arroja. Estas métricas son calculadas como el promedio simple de los resultados de las k-iteraciones que se realizan.

2.6. Algoritmos de clasificación

2.6.1. Naïve Bayes

Es un clasificador probabilístico ya que está basado en el teorema de Bayes para indicar la probabilidad de que una instancia pertenezca a cada clase. Este algoritmo asume independencia entre los atributos dada una clase.

Para efectos de la clasificación de texto, este clasificador puede utilizarse considerando que el documento se representa mediante un vector binario que indica la presencia o ausencia de cierto término. Pero, este no se puede utilizar en el caso de que se tengan las frecuencias de cada término. Para dichos casos, se puede implementar un clasificador Naïve Bayes Multinomial como una alternativa para utilizar matrices de término-frecuencia [20].

En el modelo Naïve Bayes tradicional, o también conocido como modelo Bernoulli multivariado, un documento es un vector $D = (\omega_1, \omega_2, ..., \omega_{|V|})$ de dimensión |V| siendo V el vocabulario y ω_i una variable binaria que representa la presencia o ausencia del termino i-ésimo del vocabulario.

Por otro lado, en el modelo Naïve Bayes Multinomial un documento es un secuencia ordenada de la ocurrencia de los tokens de manera independiente. Este no considera la información del orden de los tokens en el documento, sólo ocurrencia [30].

Lo que retorna tanto Bernoulli multivariado como Naïve Bayes Multinomial en minería de texto es la probabilidad de que un documento d pertenezca a la clase c. Y así, definiendo un *threshold*, el modelo clasifica por clase a cada instancia.

De las ecuaciones 2.1, 2.2 y 2.3 se observa cómo se define la clase más probable, C*, a la que pertenece un documento D utilizando el Teorema de Bayes [28]:

$$C^* = argmax P(C|D) \tag{2.1}$$

$$C^* = argmax P(D|C)P(C) \tag{2.2}$$

$$C^* = argmax P(\omega_1, \omega_2, ..., \omega_{|V|}|C)P(C) \tag{2.3}$$

Siendo ω_i una variable binaria de ocurrencia en el modelo de Bernoulli multivariado y una variable continua que representa frecuencia de ocurrencia en Naïve Bayes Multinomial.

2.6.2. Support Vector Machine

Si bien en sus inicios este clasificador se diseñó para resolver problemas de clasificación binaria, hoy en día se le han ido incorporando variaciones mediante las cuales el algoritmo se utiliza para regresiones y clasificación binaria y multiclase.

La máquina de vectores de soporte o *Support Vector Machine* (SVM) implementa la idea de mapear los atributos que se ingresan como input en un espacio multidimensional. En este espacio se construye un vector lineal que es capaz de separar a ambas clases, si se considera clasificación binaria [11]. Dicho vector lineal es el hiperplano que permite separar las clases y los vectores de soporte son aquellos puntos de cada una de las clases que están más cerca del hiperplano y que son los que definen el margen de máxima separación entre ambas clases [11].

En la Figura 2.4 se observa un ejemplo de un problema separable con SVM en un espacio de 2 dimensiones.

Sea

$$\omega_0 * z + b_0 = 0 \tag{2.4}$$

el hiperplano óptimo del espacio de los atributos, los pesos ω_0 pueden escribirse como una combinación lineal de los vectores de soporte, es decir, de la siguiente forma:

$$\omega_0 = \sum_{\text{vectores de soporte}} \alpha_\mathrm{i} z_\mathrm{i} \tag{2.5}$$

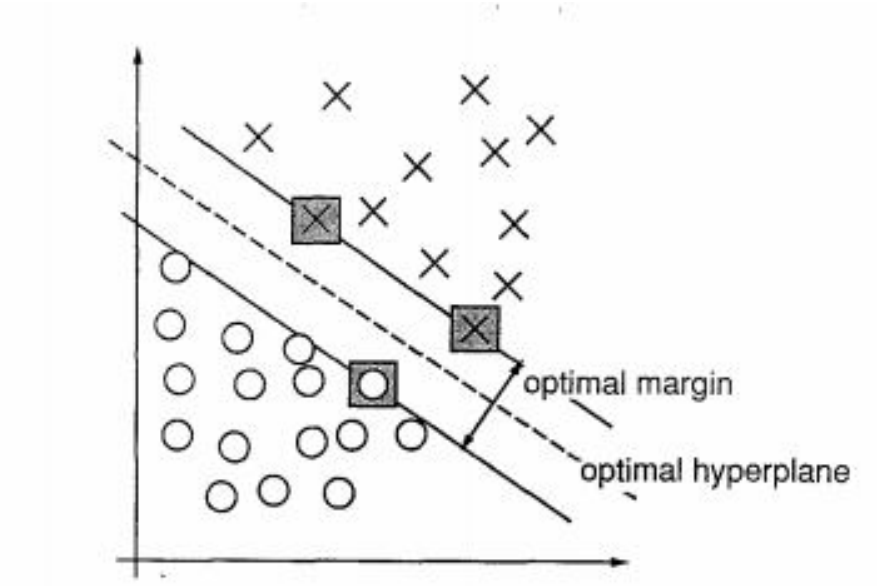

Figura 2.4: Ejemplo de un problema separable con SVM en un espacio de 2 dimensiones
Fuente: [11]

Además, se define el margen como la suma de las distancias de los puntos más cercanos al hiperplano óptimo y el objetivo matemático es maximizar dicho margen [16].

No siempre un hiperplano lineal es capaz de separar de la mejor forma posible las clases [16]. Es por esto que existen diferentes formas para poder calcular el hiperplano óptimo de manera de generalizar lo mejor posible la separación de las clases. Estas variaciones permiten construir un clasificador a partir de definir el hiperplano de diferentes formas.

Stochastic Gradient Descent (SGD)

Es un enfoque que se utiliza con clasificadores lineales para funciones de pérdida convexas, y típicamente se utiliza con SVM y regresión logística cuando se tiene una alta cantidad de datos y es útil para minería de texto por los datos poco densos que se tienen con texto [3].

Para minimizar una función mediante el gradiente se toman pequeños pasos en la dirección para llegar al mínimo, pero en el caso de una alta cantidad de datos se tendrían que calcular muchos ejemplos y esto no es óptimo. Por esto, SGD propone elegir en cada paso un vector aleatorio y calcular el gradiente de dicho vector [23].

Para lo anterior, SGD entrena un modelo eligiendo cierta cantidad de *epochs* y número de etapas para cada *epoch*. Luego, elige aleatoriamente un punto de partida y para cada epoch da un paso en dirección del gradiente del vector. Y así sucesivamente hasta encontrar el mínimo [23].

2.6.3. Random Forest

Es un algoritmo, utilizado tanto para problemas de clasificación como para regresiones, que forma diferentes árboles de decisión para luego en conjunto poder obtener un clasificador con una mejor predicción.

Tanto para la tarea de clasificación como de regresión, lo que hace Random Forest es considerar aleatoriamente ciertos atributos que se le dan al algoritmo, luego con dichos atributos se genera un árbol de decisión. Este proceso se realiza muchas veces para conjuntos de atributos aleatorios. Finalmente, según el resultado que cada árbol de decisión arroja, la categoría que tuvo mayor frecuencia es aquella que será el output de Random Forest.

El resultado que se obtiene del algoritmo proviene de la combinación de los resultados obtenidos por cada árbol individualmente. En el caso de la clasificación, el resultado de Random Forest es aquella clase que fue más popular como respuesta de cada árbol de decisión entrenado [8]. Se puede obtener la importancia de cada atributo que se considera en el input, de manera de saber cuáles son aquellas más relevantes para determinar el resultado.

Este algoritmo se construye considerando un set de atributos en donde para el k-ésimo árbol se genera un vector aleatorio Ω_k independiente de los vectores aleatorios de los árboles anteriores (Ω_1,...,Ω_{k-1}) pero con la misma distribución. Se construye el árbol k utilizando el set de entrenamiento y Ω_k, en donde resulta un clasificador $h(x,\Omega_k)$, donde x es el input. El conjunto de clasificadores $h(x,\Omega_1)$,...,$h(x,\Omega_k)$ estructurados como árbol forman el clasificador Random Forest [8].

Se recomienda ya que es un modelo que generalmente es rápido, simple y flexible [14].

2.7. Procesamiento de texto

2.7.1. Matriz Tf-Idf

Su nombre proviene de la abreviación de frecuencia de término y frecuencia inversa de documento, lo cual en inglés es 'Term frequency - Inverse document frequency'. Esta calcula la frecuencia en que ocurre cada término de una colección de documentos, pero a su vez indica la relevancia de dicha palabra dentro del documento ya que 'castiga' a aquellos términos que son muy utilizados en toda la colección de documentos, independiente de la clase.

Esta matriz es generalmente utilizada en minería de texto para transformar los términos de documentos en un valor que le permita a los algoritmos de aprendizaje automático entender a lo que se está enfrentando, ya que éstos no son capaces de 'leer' tokens.

El Tf, es decir la frecuencia de término tf(t,d) calcula la cantidad de veces que el termino t aparece en el documento d. Por otra parte, la frecuencia inversa del documento idf(t,D) mide si el termino t es común dentro de toda la colección de documentos D. Entonces tf-idf pondera ambas medidas y es por esto que se castiga a los términos que son muy comunes.

En minería de texto, aquel modelo que utiliza matriz tf-idf como input para ser entrenado se llama Modelo Bag-Of-Words.

2.8. Evaluación de resultados de modelos de clasificación

2.8.1. Matriz de confusión

Es una matriz en que se tabula el conteo de registros que se predijeron correcta e incorrectamente según cada clase. Ésta es bastante útil para realizar un análisis visual sobre el desempeño del modelo en relación a la cantidad de casos por cada clase en que se clasificó bien y si existe una tendencia a confundirla con otra clase, por lo que no sólo se analiza el error a nivel global del modelo. En el caso de la clasificación binaria, en esta matriz se consideran los registros que fueron clasificados con la misma clase de la que realmente son (Verdaderos Positivos y Verdaderos Negativos) y los registros que fueron mal clasificados (Falsos Positivos y Falsos Negativos). Los 'positivos' y 'negativos' hacen alusión a las dos clases posibles donde típicamente la clase 'positiva' se asigna a la clase de interés mientras que la 'negativa' es la restante. Un ejemplo de esto es que en un problema de clasificación en que se busca predecir fraude o no fraude, la clase 'positiva' es fraude ya que es la que es de interés predecir. En el caso de 3 o más clases, la matriz de confusión se comporta de la misma forma, considerando que los verdaderos son los que fueron correctamente clasificados.

En las columnas de la matriz se ubican cada una de las clases reales, por lo que la suma de registros por columna debe resultar en el número total de registros por cada clase. En las filas se ubican los registros que fueron clasificados por el modelo a cada clase. En la Figura 2.1 se puede observar un ejemplo de matriz de confusión para clasificación binaria.

Sea VP: Verdadero Positivo, VN: Verdader Negativo, FP: Falso Positivo y FN: Falso Negativo.

		Clase Predicha	
		Verdadero	Falso
Clase real	Verdadero	VP	FN
	Falso	FP	VN

Tabla 2.1: Matriz confusión clasificación binaria
Fuente: Elaboración propia

A partir de los valores de la matriz de confusión se pueden calcular las siguientes métricas [3]:

[3]Se asume clasificación binaria para las fórmulas de las métricas

2.8.2. Métricas de desempeño

Exactitud (*Accuracy*)

Esta métrica se calcula para analizar la capacidad de predecir correctamente todas las clases.

$$Accuracy = \frac{\text{Nº registros clasificados correctamente}}{\text{Nº total de registros}} = \frac{VP + VN}{VP + VN + FP + FN} \tag{2.6}$$

Precisión

Esta métrica se calcula para analizar la capacidad de predecir correctamente según cada clase del modelo.

$$Precision = \frac{\text{Nº registros clasificados correctamente de la clase positiva}}{\text{Nº registros clasificados correctamente}} = \frac{VP}{VP + VN} \tag{2.7}$$

Recall

$$Recall = \frac{\text{Nº registros clasificados correctamente de la clase positiva}}{\text{Nº registros que en realidad son de la clase positiva}} = \frac{VP}{VP + FN} \tag{2.8}$$

F1 Score

$$F1 - Score = \frac{2 * Recall * Precision}{Recall + Precision} \tag{2.9}$$

ROC

La curva ROC es una gráfica que representa la tasa de verdaderos positivos por sobre la tasa de falsos positivos en diferentes umbrales de clasificación.

Esta gráfica se realiza considerando el eje X (tasa de falsos positivos) entre 0 y 1 al igual que el eje Y (tasa de verdaderos positivos). Dado esto, mientras la curva sea lo más parecido a una curva logarítmica muy cercana al eje Y, mejor es el modelo.

AUC

El área bajo la curva ROC es lo que se llama la métrica AUC. Ésta calcula el área que abarca la curva ROC para obtener una medida agregada del rendimiendo de los diferentes umbrales de clasificación que se grafican con ROC.

Se interpreta el AUC como la probabilidad de que el modelo entrenado clasifique como positivo a un registro positivo aleatorio que un registro negativo aleatorio.

Se utiliza esta métrica para evaluar entre modelos ya que no varía con respecto a la escala de los valores, sólo mide qué tan bien clasifica el modelo. Pero, en el caso que se desee priorizar la detección de falsos positivos, el AUC no es una buena métrica.

Error

$$Error = \frac{\text{Nº registros clasificados incorrectamente}}{\text{Nº total de registros}} = \frac{FP + FN}{VP + VN + FP + FN} \tag{2.10}$$

2.8.3. Error de entrenamiento

Este error se calcula ya que al entrenar un modelo este puede aprender muy bien sobre dichos datos pero no generaliza, es decir,que al preguntarle sobre nuevos datos no tiene buena capacidad de predicción. En esos casos, se dice que el modelo está sobre-ajustado o sobre-entrenado (*overfitting*).

2.9. Predicción de ubicación geográfica en Twitter

En el contexto de la predicción de ubicación geográfica de usuarios en Twitter, la literatura ha estudiado 3 tipos: Home Location, Tweet Location y Mentioned Location. [33]

- Home Location se refiere a la ubicación residencial de largo plazo. Esta puede ser en base a:
  1. Regiones administrativas como lo son el país, estado, ciudad, región, etc.
  2. Grillas geográficas que están dadas por celdas de un cierto tamaño que particionan a la región que se busca estudiar. La Home Location estaría dada por la celda a la que el usuario pertenece.
  3. Coordenadas geográficas dadas por una latitud y longitud.
- *Tweet Location* se refiere a la ubicación en la que fue realizado el post que en general están dadas por coordenadas geográficas.
- *Mentioned Location* se refiere a la(s) ubicación(es) que un usuario menciona en el contenido del tweet.

Dentro del estado del arte en este tipo de predicción, se encuentran investigaciones como [19] que utiliza datos desde la API que cumplen con que en el campo 'location' saliera Qatar, el país en el cual se basó la investigación, o que al menos tuviera un *tweet* con *geotag* en Qatar. Se recolectaron 51.449 usuarios en 2 meses junto a todos sus *tweets*, que resultaron un total de 54.075.860. Utilizaron contenido del *tweet*, latitud y longitud de los *geotag* y la información sobre el dispositivo que se utilizó cuando se publicó el *tweet*. Se filtraron los usuarios inactivos exigiendo 10 tweets con al menos 5 seguidores y 5 seguidos. Obtuvieron perfiles de seguidores y seguidos de los usuarios que recopilaron antes y las fotos de perfil. Clasificación de la data mediante crowdsourcing. Una de las variables que más aporta al modelo son los hashtags y

el lenguaje del tweet. El modelo que se utilizó para predecir nacionalidad fue Gradient Boost Tree el cual resultó ser el que arrojó un mejor desempeño con un *accuracy* de 83,8 %.

Por otro lado, el modelo Naive Bayes Multinomial se encuentra como el modelo que más ha resultado en otras investigaciones relacionadas a la predicción de geolocalización, tanto a nivel de ciudad en algunos casos [17] como a nivel país o ambas [18].

En [10] utilizan dicho modelo para determinar ubicación del hogar a nivel de ciudad de los usuarios de Twitter utilizando palabras indicativas a través de *text mining* del contenido del *tweet*, lugares mencionados en el *tweet* y los hashtags que se utilizan. Este modelo, sólo utilizando el contenido del *tweet* logra un accuracy de 22,5 % en testing y 19,5 % en validación.

En [6], utilizando el mismo modelo se predice la ubicación frecuente a nivel de ciudad mediante un enfoque basado en la red y recopilando las entidades geográficas relevantes de la descripción del usuario. Dicho modelo alcanzó más del 57 % de *accuracy*.

2.9.1. Text mining para determinar *Home Location*

Se han utilizado múltiples formas para determinar Home Location de usuarios. Estas diferentes metodologías están asociadas a indicadores espaciales, como se definen en [5]. Y son:

- Ubicaciones mencionadas en el tweet
- Red de seguidores y seguidos
- Campo Location
- Dirección IP de sitio web
- Geotags del tweet
- Links URL ([31]) en que se detectan las URL del cuerpo del tweet y se asocian al país/ciudad/región/estado de la dirección
- IP del dominio del servidor.
- Zona horaria
- Metodología híbrida que utiliza más de un indicador espacial.

Según lo indicado en el estudio más reciente sobre Home Location [33], la forma de abordar esta investigación según diferentes autores ha sido la siguiente:

La mayoría de los trabajos previos se han realizado en base al contenido de los *tweets*. El método *word-centric* se basa en el cálculo de la probabilidad de estar en la ubicación u dado que se han escrito las palabras w, mientras que el método *location-centric* se basa en el cálculo de la probabilidad de escribir las palabras w dado que se está en la ubicación u, y los métodos híbridos combinan los dos anteriores. En cuanto al *dataset*, la gran mayoría ha utilizado en base a *tweets* con *geotag* y para asumir la ubicación de la persona considerándola verdadera se ha utilizado *geotags* junto con lo ingresado en el campo *location*.

2.9.2. Métricas de evaluación de desempeño para geolocalización en *Twitter*

Twitter cuenta con una gran cantidad de datos tanto por la cantidad de cuentas de usuarios que posee como los *tweets* que estos realizan. Dado esto es que los problemas de clasificación en *Twitter* no son un tema sencillo de evaluar ya que depende bastante de la muestra inicial con la que se trabaje que es difícil de estratificar según las clases que se busca predecir. A partir de esta problemática es que se han propuesto y utilizado métricas que se ajusten a este tipo de problemas y que permiten medir si el algoritmo/modelo que se utiliza para extraer usuarios o *tweets* con alguna característica en común está efectivamente recopilando lo que debería. Para medir esto se utiliza *retrieval recall* y *retrieval precision* [21].

Aquella métrica relacionada a la precisión mide cuántos datos de aquellos que se recopilaron no son desperdicio, es decir, datos que se recopilaron que cumplen con las características de aquellos que debían recopilarse. Aquella relacionada al *recall* mide cuántos datos de los que deberían haberse recopilado efectivamente se recopilaron.

2.10. Datos personales

En Chile, la regulación constitucional de la protección de datos personales está dada por la Ley N° 19.628 "Sobre protección de la vida privada y Protección de Datos de Carácter Personal"que fue promulgada en agosto del año 1999.

La protección de los datos personales se puede definir como "...el amparo debido a los ciudadanos contra la posible utilización por terceros, en forma no autorizada, de sus datos personales susceptible de tratamiento autorizado, para de esta forma, confeccionar una información que, identificable con él, afecte a su entorno personal, social o profesional, en los límites de su intimidad" [13]. En la cual se deja ver un dato personal como aquella información que se utiliza por terceros, sin autorización previa de la persona física a quien le pertenece, que la hace identificable y afecta su intimidad.

Según el Artículo 2° de la Ley antes mencionada, se explicita que se entenderá por Datos de carácter personal o datos personales como "los relativos a cualquier información concerniente a personas naturales, identificadas o identificables". Es decir datos como el nombre completo, dirección postal, teléfono, número de patente vehicular o dirección IP son ejemplos de datos que hacen identificable a una persona ya que a través de ellos se podría saber a quién pertenecen los datos si se cruzan.

En el marco de la geolocalización a nivel regional de los usuarios chilenos de Twitter se obtendrá información de los usuarios que sean utilizados en la base de entrenamiento mediante los datos públicos que se comparten en la red de *Twitter*. Dicha información es pública sí y solo sí cada persona física dueña de una cuenta de *Twitter* autoriza a *Twitter* a que su información esté disponible públicamente. Mediante dicha información, se busca obtener, con un cierto nivel de probabilidad, si dichos usuarios viven en Chile o no, y en caso de vivir en Chile, en cuál de las 15 regiones administrativas actuales viven. Dado esto, se va a intentar

predecir uno o más datos de los usuarios que no son considerados personales ya que ni el país ni la región son datos que podrían hacer identificable a una persona.

Capítulo 3

Construcción set de datos

Para el diseño y construcción de los modelos es fundamental contar con una base de entrenamiento etiquetada de manera de generar modelos supervisados que luego permitan evaluar el desempeño de clasificación. El proceso de construcción de los set de datos para ambos modelos es el que toma más tiempo ya que no es trivial contar con una base extensa de usuarios que se sepa con certeza que viven en Chile o en otro lugar del mundo, y si viven en Chile en cuál de las quince regiones viven. Es por esto que se toman una serie de supuestos para armar una base lo suficientemente grande y lo menos sesgada posible para entrenar los modelos necesarios.

En la sección 3.1 se realiza una descripción de los modelos a realizar, luego en la sección 3.2 se detalla las directrices que se tomaron para la obtención de datos, la forma en que fue realizada y las fuentes desde las cuales se obtuvieron dichos datos. A continuación, en la sección 3.3 se describe el proceso que se realizó para la selección de los datos anteriormente obtenidos para la construcción de las bases de entrenamiento de ambos modelos, explicitando las formas de etiquetado que se utilizaron. Finalmente, en la sección 3.4 se realiza un análisis descriptivo de las bases de entrenamiento.

3.1. Definición de modelos

3.1.1. Modelo de clasificación: País

Las clases a predecir del modelo país son si el usuario vive o no vive en Chile, por lo tanto el modelo sólo se enfoca en Chile y no se podrá saber a qué país pertenece un usuario si es que este no es clasificado como chileno. Para esto, se considera todo el universo de usuarios de *Twitter* para extraer una muestra de usuarios chilenos y usuarios de otros países para entrenar dicho modelo. Dado lo anterior, el modelo país se construye como un modelo binario, supervisado, con clases nominales y balanceadas.

3.1.2. Modelo de clasificación: Región

En el caso del modelo de clasificación para determinar la región a la que pertenece un usuario chileno de *Twitter*, las clases a predecir son: Región de Arica y Parinacota, Región de Tarapacá, Región de Antofagasta, Región de Atacama, Región de Coquimbo, Región de Valparaíso, Región del Libertador Bernardo O'Higgins, Región del Maule, Región del Bío Bío, Región de la Araucanía, Región de los Ríos, Región de los Lagos, Región de Aisén del General Carlos Ibáñez del Campo, Región de Magallanes y Antártica Chilena y Región Metropolitana.[1] Por lo tanto, modelo región es un modelo de clasificación multiclase, supervisado, con clases nominales y balanceadas.

3.2. Obtención de datos

Para construir los modelos es necesario contar con datos de entrenamiento lo menos sesgados posible de manera que luego los modelos reciban como input datos de nuevos usuarios y que su desempeño sea similar al de después de haber sido entrenado y calibrado. Para esto, es necesario considerar una recopilación de datos que sea suficiente tal que el modelo no sobre-ajuste y que dicha muestra sea representativa, es decir, que al momento de seleccionar a los usuarios sea una muestra estratificada aleatoria para así disminuir el sesgo de selección.

3.2.1. Recopilación de usuarios

Definir la base de entrenamiento no es tarea fácil dado que no se cuenta con usuarios etiquetados al extraer datos de *Twitter* con respecto al país que un usuario vive, ni menos de la región en que viven para el caso de usuarios chilenos. Para esto, es necesario definir una forma de obtener esta información etiquetada de manera de lograr entrenar modelos supervisados. Según lo expuesto en la sección 2.8 sobre lo realizado en otros estudios sobre la predicción de la geolocalización con datos de *Twitter*, se observa que lo más utilizado para definir la geolocalización verídica de un usuario es el campo *location* y los *tweets* geolocalizados. Si bien hay estudios que sólo utilizan el campo *location*, otros que sólo utilizan *tweets* geolocalizados y otros que utilizan ambas características; se opta por utilizar ambas características ya que de esta forma se disminuye el sesgo de selección.

Entonces, para incorporar ambas características y considerando además que se cuenta con datos de una encuesta que realizó el WIC que contiene información de cuentas de usuarios chilenos de *Twitter* junto con la región en que viven, se realiza el proceso de recopilación de usuarios. Este cuenta con tres formas de obtención de usuarios que provienen de dos fuentes de datos: API REST de *Twitter* y la encuesta.

Para extraer datos desde la API REST se utiliza un módulo de recolección de datos [12] implementado en el WIC, que utiliza una cola iterativa de credenciales para acceder a la

[1]La nueva división política y administrativa de Chile considera una nueva región, la Región del Ñuble, que entra en vigencia a fines del 2018 por lo que no se considera dentro de las clases.

API REST y así no verse afectado por la restricción de *Twitter* sobre el límite de número de peticiones cada 15 minutos por cada credencial. Este módulo de recolección utiliza un algoritmo de recorrido de grafos llamado Búsqueda de Anchura el cual para cada elemento de la red se agregan todos los elementos adyacentes a él que cumplan con cierta condición. En este caso, un conjunto de usuarios son los elementos de la red y los elementos adyacentes a él son aquellos usuarios que siguen a dicho elemento. Las condiciones que deben cumplir para ser agregados a la cola son que su campo location debe contener la palabra 'Chile' o bien cumplir con que su atributo 'isgeoenabled' sea igual a 'true'.

Para lo anterior, se utilizaron cincuenta semillas de usuarios chilenos y cincuenta semillas de usuarios extranjeros, para así asegurarse de abarcar usuarios de todo el mundo pero también que se capturen varios chilenos. Las semillas se usuarios chilenos se obtuvieron tras la recopilación de datos durante una semana desde la API Streaming de *Twitter* utilizando el *bounding box* al que pertenece Chile [2]. Se utilizan aquellas cuentas que tenían más seguidores al momento de recopilar los datos y que pertenecían a Chile. Cuentas como 'Ozzy Osbourne' que se detectó en Chile dado que publicó un *tweet* diciendo que estaba Chile en el periodo que fueron recopilados los datos por un próximo concierto que haría en el país, no se consideraron ya que no vive en Chile. Todas las cuentas identificadas se validaron manualmente. Entre ellas se encuentran las cuentas de: Sebastián Piñera, Radio ADN, Carabineros de Chile, Centro GAM, Amaro Gómez Pablo, Francisco Saavedra, Claudio Orrego, Sismos en Chile, entre otras. También se utilizaron 50 semillas extranjeras que se obtuvieron tras la recopilación de *tweets* durante 24 horas [3] los cuales se ordenaron por país y se eligió el usuario con más seguidores de cada uno de los 50 países con más *tweets*.

Tras conocer la metodología de extracción mediante un *crawler* desde la API REST de *Twitter*, a continuación se detallan las tres formas de obtención mediante las cuales se extrajeron ciertas cuentas de usuarios de *Twitter* para luego construir las bases de entrenamiento con usuarios chilenos y extranjeros:

1. **Campo 'isgeoenabled' = 'true':**
   Cuentas de usuarios recopiladas mediante un *crawler* desde la API REST de *Twitter* que cumplan con que el campo 'geo_enabled' sea igual a 'true', ya que esto permite obtener sólo aquellos usuarios cuyos *tweets* podrían tener el campo 'place' y por ende se podría determinar el país en que se publicó. [4]. Se obtienen 253.383 usuarios mediante esta forma.

2. **Campo '*location*' contiene la palabra 'chile':**
   Cuentas de usuarios recopiladas mediante un *crawler* desde la API REST de *Twitter* que almacena sólo aquellos usuarios que cumplen con que aparezca 'Chile' en el campo *location* de su perfil de usuario. Éstos fueron recolectados desde el Viernes 20 de Abril 2018 hasta el Viernes 4 de Mayo de 2018.

   Con esta modalidad se logran extraer 72 usuarios que según su perfil son de Chile dado que aparece Chile en su campo *location*. Se asume que todos dichos usuarios

[2]Coordenadas Suroeste: (-75.56, -55.67) y coordenadas Noreste (-67.73, -17.15)

[3]Desde el miércoles 18 de abril 2018 a las 13:06 hrs hasta el jueves 19 de abril 2018 a las 13:33 hrs

[4]Sólo tendrán el campo 'place' no nulo aquellos *tweets* que hayan sido publicado después de que el usuario haya activado la geolocalización y que el usuario haya puesto un lugar asociado al *tweet*.

son efectivamente chilenos para efectos del modelo a realizar. El *crawler* es capaz de recolectar pocos usuarios que cumplen con la restricción dado que las semillas utilizadas son de usuarios de todo el mundo.

Para aumentar la cantidad de usuarios de esta forma de obtención, se filtra la base de chilenos que actualmente tiene el WIC en base a la misma restricción, es decir, que contenga la palabra 'chile' en su campo *location*. Luego de este filtro se obtienen 1.675.850 cuentas de usuarios, y sumado a las obtenidas por el *crawler* se llega a un total de 1.675.922 ya que no habían cuentas duplicadas.

3. **Encuesta en Twitter:**
   Usuarios chilenos de *Twitter* que contestaron una encuesta realizada por el WIC, en el contexto del proyecto SONAMA, indicando sus datos personales, incluyendo región en la que vive y su cuenta de *Twitter*. El total de usuarios asociados a esta forma de obtención es de 634.

Dado lo anterior, inicialmente se cuenta con un total de 1.929.939 cuentas de usuarios según lo indicado en la tabla 3.1.

Procedencia	Cantidad de usuarios
Geolocalizados	253.383
Location 'chile'	1.675.922
Encuesta	634
Total:	**1.929.939**

Tabla 3.1: Base total de usuarios de *Twitter* obtenidos y su procedencia

Los usuarios extranjeros se obtienen mediante la base de usuarios con geolocalización activada mientras que los usuarios chilenos se obtienen mediante las formas indicadas anteriormente. La figura 3.1 detalla las fuentes de las cuales se extraen los usuarios para construir las bases de entrenamiento.

3.2.2. Recopilación de los *tweets* de los usuarios obtenidos

Para cada usuario obtenido se determina que se recopilarán los *tweets* del último año móvil con un máximo de 200 *tweets*. Esto ya que los datos de un año son más estables para determinar dónde se estuvo con mayor frecuencia para determinar donde vive actualmente la persona y la cota máxima se introduce para manejar mejor los datos ya que al considerar la base de entrenamiento de mínimo 15.000 usuarios y cada uno con 200 *tweets* máximo, es un número razonable para procesar. Los *tweets* que se extrajeron por cada usuarios fueron los emitidos entre el 2 de Mayo de 2017 y 12 de Mayo de 2018.

Estos fueron recopilados desde la API REST de *Twitter* mediante un *crawler* que recorre cada ID de usuario que se había obtenido previamente en la base de geolocalizados, base *location* y base encuesta. Por cada usuario captura los *tweets* públicos emitidos entre el rango de tiempo estipulado.

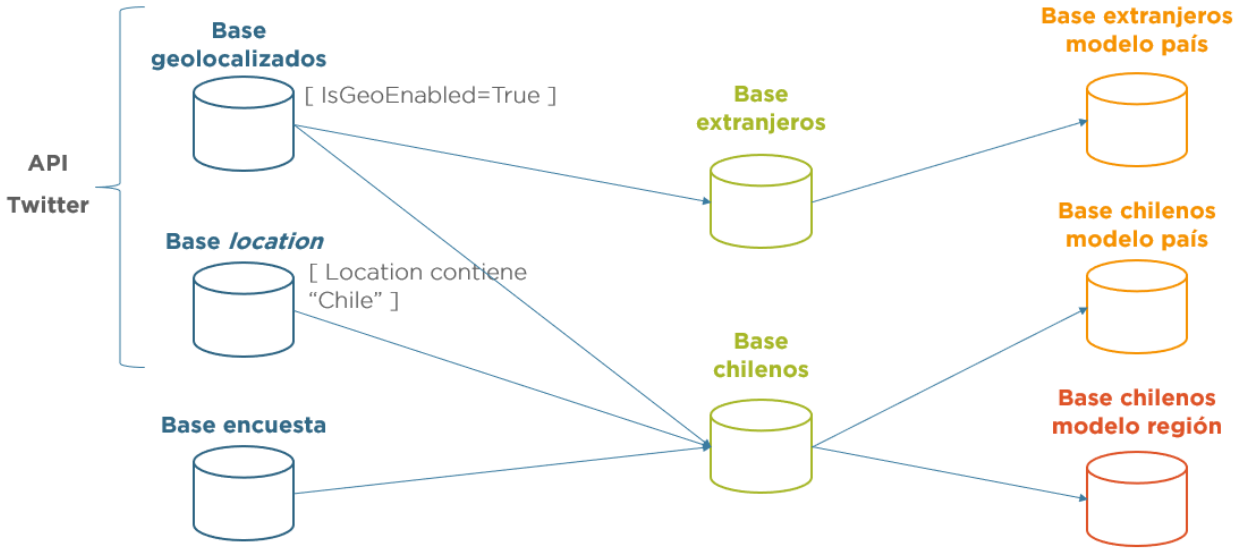

Figura 3.1: Fuentes de datos de bases de entrenamiento
Fuente: Elaboración Propia

Cada *tweet* contiene los atributos [5] mencionados a continuación en la tabla 3.2.

Campo	Descripción
twitter_id	ID del *tweet*
user_id	ID del usuario que emitió el *tweet*
text	texto de la publicación (puede ser *tweet* o *retweet*)
createdat	fecha y hora en que se realizó el *tweet*
lang	lenguaje en que se realizó el *tweet*
geo	JSON con atributos de la geolocalización automática del *tweet*
place	JSON con atributos de la geolocalización que el usuario incorpora al *tweet*

Tabla 3.2: Campos de atributos extraídos de cada *tweet*

3.2.3. Etiquetado

Luego de realizar la obtención de usuarios chilenos y extranjeros mediante diferentes métodos y obtener sus respectivos *tweets*, se procede a etiquetar a los usuarios de manera de obtener cuáles son chilenos y cuáles extranjeros, y de los chilenos se debe obtener la región en la que viven.

Dependiendo de la base de procedencia del usuario es el atributo que se le considerará para ser etiquetado. Esto es:

- Base de procedencia **campo location**: Se utiliza información del campo *location* para ser etiquetada.

[5] Cabe destacar que dichos atributos son sólo los que se decidieron extraer, existen más campos que describen cada *tweet*.

- Base de procedencia **geolocalizados**: Se utiliza información del campo *'place'* [6] de todos los *tweets* que tengan dicho campo no nulo. En particular, se utiliza el atributo *'country'* y atributo *'name'*.
- Base de procedencia **encuesta**: Se utiliza información externa de *Twitter*. Son datos que posee el WIC sobre algunos usuarios y se conoce si es que actualmente ese usuario vive en Chile y en qué región.

Etiqueta país

Para etiquetar el país en el cual vive cada usuario recopilado, se realiza lo siguiente:

- **Base encuesta**
  La encuesta fue realizada a usuarios chilenos de *Twitter*, por lo que se etiquetan a todos como chilenos excepto aquellos que respondieron que viven fuera de Chile. Por lo tanto, todo usuario que es parte de la base de encuesta que haya respondido que vive en una de las quince regiones administrativas de Chile es agregado a una tabla auxiliar que contiene las columnas de ID de usuario, su *screename* y una columna país. Estos se etiquetan como 'Chile' en la columna 'pais' de la tabla.

- **Base location**
  En el caso de los usuarios que provienen de la base de aquellos que contienen 'chile' en su campo *location*, según los supuestos que se toman en la investigación se asume que todos estos usuarios viven actualmente en Chile. Es por esto que se realiza el mismo procedimiento anterior y todos se agregarn a una tabla auxiliar que posee las columnas 'user_id', 'screename' y 'pais', y en la columna de país todos se etiquetan como 'Chile'.

- **Base geolocalizados**
  Las etiquetas de esta base son las que más tiempo toman ya que la etiqueta está basada en el conjunto del campo *place* de cada *tweet* de cada usuario y no en un atributo del usuario de por sí, como lo es el campo *location*. Por lo tanto, para definir la etiqueta de cada usuario se debe realizar un compilado de todos los países que fueron nombrados en los últimos 200 *tweets* de cada usuario, y aquel que haya sido nombrado con una mayor frecuencia es el país al cual se asignará el usuario.

  Para esto se realizan los siguientes pasos:

  1. Tener en la base los últimos 200 *tweets* de cada usuario.
  2. Extraer de cada *tweet* el país mencionado en el campo *place* mediante la variable 'country' del JSON del campo 'place'.

     No necesariamente todos los *tweets* de usuarios con la geolocalización activada tienen el campo *place* no nulo. Tener dicha opción activada sólo da la opción de agregar un geotag al *tweet* mas no siempre el usuario lo agrega. De hecho, un usuario con geolocalización activada podría no tener ningún *tweet* con campo *place* no nulo.

[6] Al recopilar los *tweets* de cada usuario el campo *'place'* se obtiene en formato JSON ya que dicho objeto de por sí tiene atributos contenidos, como los que se mencionan en la sección 2.3.1 en la figura 2.1.

user_id	place ->'country'
1234	Rusia
56789	null
...	...
1234	Rusia
32048	Uruguay

Tabla 3.3: Ejemplo tabla del paso 2

3. Realizar un conteo de registros agregados por el ID de usuario y el país mencionado. En este paso se excluyen aquellos *tweets* que hayan tenido el campo *place* nulo.

user_id	country	conteo
1234	Rusia	20
32048	Uruguay	8
...	...	...
1234	Islandia	2

Tabla 3.4: Ejemplo tabla del paso 3

4. Para cada usuario se elige el país en que más *tweets* tenía, es decir aquel país cuya variable 'conteo' sea la mayor.

user_id	country	conteo
1234	Rusia	20
32048	USA	32
...	...	...
593	Chile	14

Tabla 3.5: Ejemplo tabla del paso 4

5. Como en la tabla del paso anterior queda un registro por cada *user_id*, entonces el país de la columna 'country' es en el cual se asume que el usuario vive actualmente. Pero, como el modelo país tiene 2 clases: si el usuario vive en Chile o no, entonces se añade una columna 'país' a la tabla en la cual a todos aquellos usuarios que en la columna 'country' tengan a Chile, se les etiqueta como 'Chile' en la columna 'país'. A todos los usuarios que tengan un país que no es Chile dentro de la columna 'country' se les etiqueta como 'Otro' en la columna 'país'.

user_id	country	pais
1234	Rusia	Otro
32048	USA	Otro
...	...	...
593	Chile	Chile

Tabla 3.6: Ejemplo tabla del paso 5

Tras el proceso anterior, de los 253.383 usuarios de la base sólo 19.926 tenían al menos un *tweet* con campo *place* no nulo. Es decir, al finalizar el paso 5 resultaron esa cantidad de usuarios etiquetados. De esos 19.926 usuarios, la gran mayoría eran de la categoría 'Otro'. La clase 'Otro' considera 184 países diferentes del mundo. La tabla 3.7 resume los números finales de etiquetado.

Etiqueta	**Cantidad**
Chile	67
Otro	19.859
Total	**19.926**

Tabla 3.7: Cantidad de etiquetados por clase para la base de geolocalizados

Resumiendo el proceso de etiquetado por país, de las tres bases resultaron las cantidades observadas en la tabla 3.8.

Base	**Nº etiquetados como 'Chile'**	**Nº etiquetados como 'Otro'**
Encuesta	618	0
Location	1.675.922	0
Geolocalizados	67	19.859
Total	**1.676.607**	**19.859**

Tabla 3.8: Etiquetados por cada clase modelo país

Etiqueta región

Para etiquetar la región de Chile en que los usuarios viven, se utilizan sólo aquellos en que en el proceso de etiquetado de país fueron identificados como chilenos. Este proceso también se subdivide según la procedencia del usuario, es decir, según la razón por la que dicho usuario fue incorporado a la base: encuesta, location contiene 'chile' o geolocalizados. Cabe destacar que si un usuario es parte de la base de encuesta, podría tener la geolocalización activada como también podría tener la palabra 'chile' en su campo *location*.

Para el etiquetado se consideraron los siguientes números de la tabla 3.9 para representar cada una de las quince regiones:

Número	Región
1	Región de Tarapacá
2	Región de Antofagasta
3	Región de Atacama
4	Región de Coquimbo
5	Región de Valparaíso
6	Región del Libertador Bernardo O'Higgins
7	Región del Maule
8	Región del Bío-Bío
9	Región de la Araucanía
10	Región de los Lagos
11	Región de Aysén
12	Región de Magallanes y la Antártica Chilena
13	Región Metropolitana
14	Región de los Ríos
15	Región de Arica y Parinacota

Tabla 3.9: Regiones con sus números respectivos de etiquetado

- **Base encuesta**
  En este caso, cada usuario declaró en la encuesta en qué región vivía actualmente por lo que se utiliza lo declarado para etiquetar en una columna llamada 'región' el número de la región.

- **Base geolocalizados + usuarios location con geolocalización**
  En el caso de los usuarios geolocalizados se realizó un proceso análogo al etiquetado de país con la diferencia que el campo 'place' entrega por sí solo un atributo 'country', mientras que para un lugar más granular sólo se tiene el campo 'name' que indica el nombre del lugar que se está incorporando. Entonces, en base al campo 'name' se identificaron todos los lugares que los usuarios chilenos etiquetaron y cada uno se asoció a una región. Luego de asociarlos a una región se realizó un conteo de veces que cada región fue nombrada por cada usuario y aquella más frecuente es la que se asume como la región en que actualmente vive el usuario. Este etiquetado se ve representado en la columna 'Geolocalizados' de la Tabla 3.10.

  En cuanto a la columna 'Geolocalizados 2', esta se construye a partir de los usuarios que inicialmente ingresaron a la base por campo 'location' pero dada la gran magnitud de esta en comparación a las demás, se identificó cuáles de dichos usuarios aparte de tener 'Chile' en su campo location, tenían la geolocalización activada y al menos 2 *tweets* con campo *place* para así etiquetar su región en base a los *tweets* más recientes y no en base a lo que indica su campo *location*. Esto último ya que se conoce que el campo *location* no es modificado muchas veces por los usuarios a pesar de cambiarse de lugar de residencia por lo que a priori la frecuencia de geolocalización de los *tweets* más frecuentes es más confiable. Entonces, a partir de dicho proceso de etiquetado se forma 'Geolocalizados 2'.

- **Base location**
  Para esta base se etiquetó por región según los nombres de región, provincia, ciudad

y/o comuna que se declaran en el campo *location*. Es decir, si en dicho campo aparece "Vivo en Maipú, la mejor comuna!", entonces dicho usuario se etiqueta como '13' (Región Metropolitana). Este proceso se realizó automáticamente a través de consultas SQL que al identificar las palabras dadas, etiquetaba al usuario según correspondiese. Se excluyeron de este proceso aquellos usuarios de la base location que fueron etiquetados mediante la metodología de etiquetado de la base de geolocalizados.

En la tabla 3.10 se observa la cantidad de usuarios etiquetados por cada una de las tres metodologías explicadas anteriormente. La región de Aysén ('11') es la región con menos usuarios identificados, lo que era esperable ya que dicha región es la que tiene una menor cantidad de habitantes según el Censo 2017 [7].

Nº región	**Encuesta**	**Location**	**Geolocalizados**	**Geolocalizados 2**	**Total**
1	4	21.663	435	1.245	**23.347**
2	18	43.658	914	2.279	**46.869**
3	8	6.973	287	667	**7.935**
4	19	41.127	969	2.299	**44.414**
5	61	91.157	3.038	7.556	**101.812**
6	21	32.406	800	2.381	**35.608**
7	20	39.928	869	1.999	**42.816**
8	82	67.052	1.990	4.540	**73.664**
9	13	36.442	807	1.859	**39.121**
10	22	37.930	869	2.357	**41.178**
11	0	4.564	150	465	**5.179**
12	6	11.854	297	781	**12.938**
13	333	483.891	13.465	33.406	**531.095**
14	8	18.332	455	1.020	**19.815**
15	3	17.195	289	470	**17.957**
Total	**618**	**954.172**	**25.634**	**63.324**	**1.043.748**

Tabla 3.10: Regiones con sus cantidades respectivas de etiquetado

3.3. Selección bases de entrenamiento

Para el modelo país es necesario obtener una muestra de usuarios extranjeros que represente el 50 % de la base total de usuarios y que sean representativos de todo el mundo, es decir que se considere una gran cantidad de países del mundo de manera de no sesgar la muestra al comportamiento principal de algún país. Dado esto, se realiza el siguiente procedimiento para definir la base de entrenamiento del modelo país:

[7] www.censo2017.cl

1. **Base de usuarios chilenos**

   (a) **Modelo país**

   Según lo expuesto en la Tabla 3.8, se cuenta con una cantidad total de 1.676.607 usuarios identificados como chilenos. De esos, según el proceso de etiquetado por región, inicialmente se lograron identificar como chilenos a través de la metodología de geolocalizados a 2.437 usuario. Esa cantidad, sumada a los 618 usuarios de la encuesta resulta un total de 3.055. A estos, se agrega una cantidad de 3.500 usuarios aproximadamente de la base *location* para equilibrar, de cierta forma, las procedencias de la base.

   Dado lo anterior, la base de entrenamiento para el modelo país cuenta con 6.555 usuarios chilenos. Cabe destacar que para efectos del entrenamiento del modelo, a dichos usuarios se les elimina la palabra 'Chile' de su campo *location* ya que eso fue utilizado para etiquetarlos. No eliminar dicha palabra podría generar sobreajuste a la base de entrenamiento.

   (b) **Modelo región**

   Considerando que el campo *location* es uno de los atributos utilizados para etiquetar la región en la que viven los usuarios chilenos, no es muy correcto que este se considere como una *feature* del modelo región ya que podría sesgar los resultados de entrenamiento. Por otro lado, otras *features* asociadas a los usuarios, sus *tweets* o la metadata como zona horaria o lenguaje del usuario no son muy representativas dependiendo de la región ya que todo Chile Continental posee la misma zona horaria y el idioma prevalente de todas las regiones es el español.

   Es por lo anterior que para el modelo región sólo se busca construir un modelo que se base en lo que publican los usuarios (*tweets* o *retweets*), por lo que es necesario contar con una base de entrenamiento de usuarios que hayan *tweeteado* al menos una vez en el periodo que se extrajeron los datos [8]. A estos usuarios se le llama usuarios activos. Aquellos que no hayan *tweeteado* ni *retweetado* dentro de dicho periodo, se les considerará como usuarios inactivos.

   Entonces, de lo expuesto anteriormente en la Tabla 3.10, se debe verificar cuántos de dichos usuarios etiquetados por región son activos para así sólo considerar estos últimos para la base de entrenamiento del modelo región.

   Al recopilar sólo aquellos usuarios activos de la base con la que se cuenta, se llegan a las siguientes cantidades expuestas en la Tabla 3.11. Sólo se modifican las cantidades de los usuarios de encuesta y los de la base location ya que los geolocalizados por construcción tienen al menos un *tweet* en el periodo en cuestión.

   La región de Aysén es la que tiene una menor cantidad de usuarios identificados con una suma de 2.286, de los cuales el 73 % corresponde a usuarios que provienen de la base location. Si se consideran clases balanceadas y fuera la misma cantidad

[8] Entre el 12/5/2017 y 12/5/2018

Nº región	Encuesta	Location	Geolocalizados	Total
1	3	7.566	1.680	**9.249**
2	8	16.255	3.193	**19.456**
3	4	1.863	954	**2.821**
4	13	15.482	3.268	**18.883**
5	46	37.376	10.594	**48.016**
6	17	12.149	3.181	**15.347**
7	12	13.329	2.868	**16.209**
8	65	25.917	6.530	**32.512**
9	9	12.935	2.666	**15.610**
10	20	13.701	3.226	**16.947**
11	0	1.671	615	**2.286**
12	4	4.473	1.078	**5.555**
13	237	193.356	46.871	**240.464**
14	8	6.695	1.475	**8.178**
15	3	5.475	759	**6.237**
Total	**449**	**368.243**	**88.958**	**457.650**

Tabla 3.11: Usuarios chilenos activos por región

para todas las regiones, la Región Metropolitana que es la que tiene una mayor cantidad de habitantes en Chile, quedaría subrepresentada.

Es por esto que se construye la base entrenamiento incorporando todos los usuarios que estén considerados de la base encuesta y se equilibra medianamente la cantidad de usuarios de la base geolocalizados con la base location según cada región. Dado esto, se incorporan más usuarios a las regiones más pobladas de manera de evitar la subrepresentación. Esto genera que las clases del modelo región estén levemente desbalanceadas pero con más información para aquellas clases que tienen más habitantes actualmente.

Tras esto, se obtiene una base de entrenamiento de 27.514 usuarios de los cuales 15.639 provienen del campo location, 11.426 de los geolocalizados y 449 de la encuesta.

La distribución por región de la base de entrenamiento es la que se observa en la Tabla 3.12.

2. **Base de usuarios extranjeros**
   (a) Esta base se construye sólo a partir de la base de geolocalizados. Según los pasos explicados previamente, de esta se obtuvieron 19.859 usuarios extranjeros correspondientes a 184 países diferentes.
   (b) De los 19.859 usuarios se eliminan 25 ya que estaban etiquetados con un país vacío, por lo que finalmente quedan 19.834 extranjeros. En la tabla 3.13 se indican los países que representan el 92 % de la base junto con las cantidades etiquetadas por cada país, mientras que en la Figura 3.2 se observa la distribución geográfica de

Nº región	**Total**
1	1.705
2	1.803
3	1.701
4	1.801
5	2.201
6	1.704
7	1.698
8	1.997
9	1.701
10	1.708
11	1.702
12	1.698
13	2.696
14	1.701
15	1.698
Total	**27.514**

Tabla 3.12: Cantidad de usuarios por región del set de entrenamiento del modelo región

dichos países en el mundo.

Dada la gran cantidad de usuarios de Estados Unidos, se realiza un muestreo aleatorio de 1.800 usuarios de dichos país de los 8.013 que hay, de manera que se acerque al número de usuarios de Reino Unido y así los usuarios extranjeros no estén sobre representados por usuarios estadounidenses.

(c) Tras lo anterior, resultan 13.621 usuarios extranjeros de la base de geolocalizados para ser utilizados para el entrenamiento del modelo país.

Resumiendo, para el modelo país se obtienen 6.555 usuarios chilenos y 13.621 usuarios extranjeros, pero se busca tener clases balanceadas por lo que se realiza un muestreo aleatorio estratificado por país de 6.600 usuarios extranjeros de los 13.621 que hay. De esto, se construye la base de entrenamiento del modelo país con 6.555 usuarios chilenos y 6.600 usuarios extranjeros.

País	Nº usuarios	País	Nº usuarios
Estados Unidos	8.013	Filipinas	154
Reino Unido	1.764	Colombia	138
Nigeria	1.401	Kenia	137
Argentina	1.246	Alemania	134
Brasil	933	Turquía	133
India	748	Venezuela	113
Canadá	582	Indonesia	101
South Africa	481	Emiratos Árabes Unidos	93
México	386	Irlanda	79
Francia	352	Ghana	79
España	347	Grecia	71
Australia	190	Chile	67
Japón	167	Holanda	62
Italia	157	Bangladesh	60

Tabla 3.13: Países y cantidad de usuarios obtenidos del 92 % de la base de geolocalizados

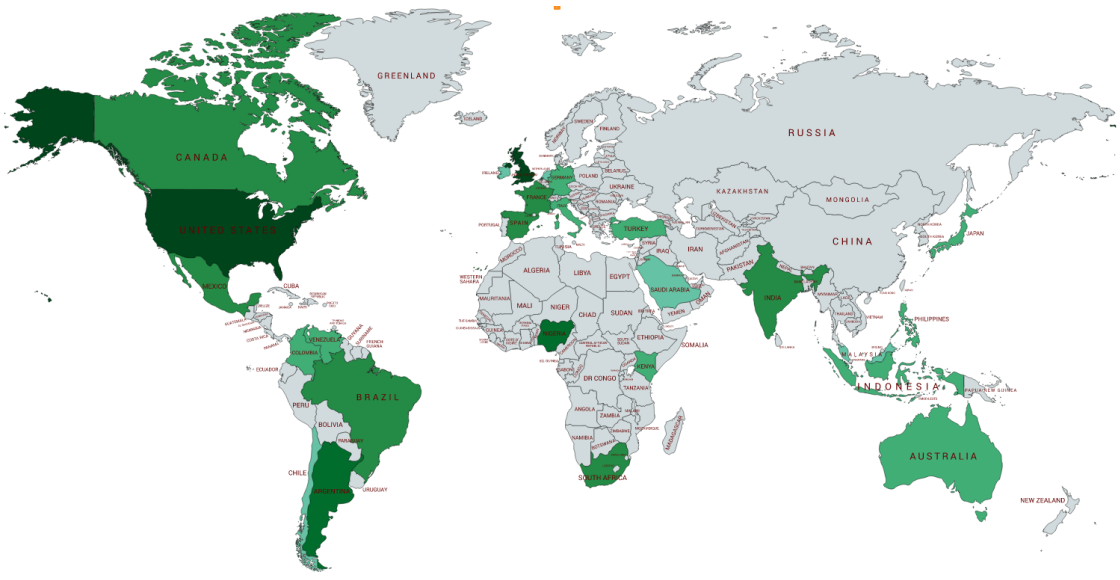

Figura 3.2: Distribución de países que representan el 92 % de la muestra de extranjeros

Capítulo 4

Modelamiento

Tras la construcción de las bases de entrenamiento, se procede a realizar el pre procesamiento de los datos y la generación de las *features* que se desean incorporar a los modelos. Luego, se entrenan varios modelos mediante diferentes técnicas de *Machine Learning*. Se probará con 3 algoritmos diferentes como lo son Naïve Bayes Multinomial, Support Vector Machine en su variación llamada Stochastic Gradient Descent (SGD) y Random Forest. La elección de dichos algoritmos se realiza en base a que éstos son los que mejores resultados han reportado en las investigaciones previas, tal como se comenta en la Sección 2.9. Ya que estos algoritmos funcionan bien con modelos multi-clases y que consideran *features* de minería de texto.

Para cada modelo se reportan los resultados obtenidos en base a las métricas de evaluación de desempeño de clasificación y matriz de confusión.

4.1. Recursos utilizados

Para entrenar los modelos que se señalan más adelante se utiliza el lenguaje de programación Python 2.7.15 mediante el entorno interactivo web Jupyter Notebook versión 4.4.0.

En cuanto a las librerías que ofrece Python, para trabajar los datos e implementar los algoritmos de Machine Learning se utiliza Sklearn.

4.2. Pre-procesamiento de datos

Para testear la hipótesis de investigación planteada se propone diseñar y construir modelos de clasificación que permitan predecir si un usuario de *Twitter* está viviendo en Chile o no, y si vive en Chile en cuál región está viviendo. Todo lo anterior, enfocado en la información del contenido y contexto de *Twitter*, es decir *tweets* y atributos que dan contexto del usuario que se pueden obtener del JSON, como:

- **Nombre (*name*)**: Nombre del usuario.
- **Nombre en Twitter (*screename*)**: El nombre del usuario que aparece en pantalla, el 'alias'.
- **Ubicación (*location*)**: Puede ser nulo. La ubicación que ingresa el usuario.
- **URL**: Puede ser nula. La URL que ingresa el usuario y que se asocia a su perfil.
- **Descripción (*description*)**: Puede ser nula. String que agrega el usuario para describir su cuenta.
- **Protegido (*protected*)**: Verdadero cuando el usuario elige proteger sus *tweets*.
- **Verificado (*verified*)**: Verdadero cuando el usuario ha verificado su cuenta.
- **Cantidad de seguidores (*followers_count*)**: Número de seguidores que la cuenta tiene.
- **Cantidad de amigos (*friends_count*)**: Número de usuarios que la cuenta sigue, es decir, número de seguidores.
- **Cantidad de listas (*listed_count*)**: Número de listas públicas que el usuario forma parte.
- **Cantidad de favoritos (*favourites_count*)**: Número de *tweets* que el usuario le ha puesto 'me gusta'.
- **Cantidad de estados (*statuses_count*)**: Número de *tweets* y *retweets* emitidos por el usuario.
- **Creación (*created_at*)**: Fecha y hora en formato UTC en que el usuario creo su cuenta de *Twitter*.
- **UTC-Offset (*utc_offset*)**: Diferencia de horario entre la ubicación del usuario y el UTC.
- **Zona horaria (*time_zone*)**: Zona horaria en que se detecta el usuario. [1]
- **Geo habilitado (*geo_enabled*)**: Verdadero cuando el usuario habilita la opción de agregarle *geotag* a sus *tweets*.
- **Lenguaje (*lang*)**: El lenguaje de la interfaz del usuario que el mismo declara.

De los atributos anteriores, algunos son características declaradas por el mismo usuario que son parte de su perfil, mientras que hay otras características que son metadata del usuario en *Twitter*.

Para efectos de los modelos a diseñar, se utilizan aquellos atributos de contexto que podrían explicar de cierta forma la ubicación frecuente actual del usuario, estos son: '*location*', 'zona horaria' y 'lenguaje'. Dichos atributos son también los que se han utilizado en investigaciones anteriores relacionadas al tema y que fueron mencionadas en la Sección 2.9.

Entonces, finalmente los atributos a utilizar en el modelo país son: 'location', 'leng', 'timezone' y conjunto de *tweets* y *retweets* realizados por el usuario entre el 12/5/2017 y el 12/5/2018. En la Tabla 4.1 se observan ejemplos de los atributos de *Twitter* que se utilizan.

Para el modelo región sólo se utiliza conjunto de *tweets* y *retweets* realizados por el usuario entre el 12/5/2017 y 12/5/2018. Se deja fuera el campo 'location' ya que si bien podría ser

[1] Desde el 23 de Mayo 2018 la zona horaria y UTC-Offset son nulos para todos los usuarios en el JSON que entrega la API REST

Atributo	Tipo	Ejemplo
Location	text	'California dreams'
Leng	string	'en'
Timezone	string	'pacific time (us & canada)'
Conjunto de tweets y retweets	text	'Donald Trump again...:(, i'm happy today '

Tabla 4.1: Ejemplo por tipo de atributo utilizado en modelo país

de utilidad para la clasificación de regiones por usuario, como dicho campo se utiliza para etiquetar la región en la base de entrenamiento, si se eliminan todas las palabras en las que se basó el proceso de etiquetado entonces el campo 'location' queda nulo o con palabras poco significativas para una gran cantidad de usuarios de la base, y por ende no contribuye al proceso de clasificación. Por otro lado, la 'timezone' para el caso de Chile no es explicativa dado que todas las regiones se encuentran en la misma zona horaria, exceptuando la Isla de Pascua que es parte de la Región de Valparaíso. Por esto, no se incluye dicha variable. Y finalmente, 'leng' tampoco es un factor diferenciador entre regiones del mismo país por lo que tampoco se utiliza.

Atributo	Tipo	Ejemplo
Conjunto de tweets y retweets	text	'todos a votar hoy!!!!'

Tabla 4.2: Ejemplo por tipo de atributo utilizado en modelo región

Para poder entrenar los diferentes modelos mediante los algoritmos de *Machine Learning* electos según las investigaciones de la Sección 2.9, lo primero que se debe realizar es una transformación de estructura de los datos para poder representar el/los atributo(s) de texto que se ingresarán a los modelos en una estructura que los algoritmos que se utilizarán puedan interpretar, y lo mismo para los atributos categóricos. Para esto, se realiza lo siguiente:

4.2.1. Atributos de texto

Tanto el atributo 'location' como el atributo 'conjunto de tw y rt' son de texto, por lo que para que los algoritmos de *Machine Learning* sean capaces de entender la información de dicho atributo es necesario realizar un procesamiento de texto que permita entregarle a los algoritmos información que puedan identificar correctamente. Para esto, se realiza la siguiente metodología:

1. **'Conjunto de *tweets* y *retweets*'**
   (a) **Eliminar URL's**:
   Se crea una función en *Python* llamada 'remove_url' para eliminar todos los elementos del texto que comienzan con 'http'.

Ejemplo *tweet* inicial:

'El día está MUY soleado... love these days with @juan123 :-) :-) #sol #playa #verano http://t.co/4389h04r'

se transforma en:

'El día está MUY soleado... love these days with @juan123 :-) :-) #sol #playa #verano'

(b) **Eliminar signos de puntuación:**
Posteriormente, se eliminan todos los signos de puntuación dados por la función *string.punctuation* de la librería *string* de *Python*.

Tweet se transforma en:

'El día está MUY soleado love these days with juan123 sol playa verano'

Es decir, en esta parte del proceso se eliminan los signos de las menciones y de los hashtags pero manteniendo el texto que acompañaba dichos signos.

(c) **Minúsculas:**
Se transforma todo el texto a minúscula para poder realizar comparación de tokens más adelante y no haya palabras iguales que no se consideren iguales por diferencias de mayúsculas.

Tweet se transforma en:

'el día está muy soleado love these days with juan123 sol playa verano'

(d) **Eliminar *stopwords***:
Dado que los lenguajes más utilizados en los *tweets* son inglés y español, entonces se eliminan las *stopwords* de ambos lenguajes.

Tweet se transforma en:

'día soleado love days juan123 sol playa verano'

(e) **Tokenización**
Se tokeniza el texto a palabras de manera que se pueda identificar explícitamente las palabras que forman el texto y así tener el *input* para realizar *text mining*. Se tokeniza separando en base a espacios en blanco.

Tweet se transforma en:

['día', 'soleado', 'love', 'days', 'juan123', 'sol', 'playa', 'verano']

(f) **Tf-Idf**:
A partir de la ponderación Tf-Idf explicada en la Sección 2.7.1, se calcula el peso de cada token, en este caso de cada palabra, del conjunto de *tweets* y *retweets* por usuario. De esta forma se obtiene una matriz *sparse* que se puede utilizar como input para la aplicación de los algoritmos de *Machine Learning*. Correr un modelo a partir de la matriz Tf-idf es lo que se llama Modelo Bag-Of-Words el cual sólo identifica los tokens en el texto con su peso y no el orden en que dichos tokens aparecen, por ejemplo.

2. **'Location'**:
En el caso de este atributo, se realiza el mismo procedimiento anterior exceptuando la eliminación de URL. Se obtiene al final una matriz Tf-idf de cada palabra con su peso.

4.2.2. Atributos categóricos

1. **Lenguaje**:

Para el atributo de lenguaje se realiza una priorización de los lenguajes a identificar para no incluir los 36 tipos de lenguajes que hay. Para esto, se observa que el 97 % de los usuarios utiliza 5 lenguajes por lo que los 31 lenguajes restantes se dejan en una categoría llamada 'otro'. Dado esto, queda un total de 6 categorías de lenguaje, que son las siguientes:

Lenguaje	Descripción	Cantidad
es	Español	7.155
en	Inglés	4.792
pt	Portugués	445
otro	Otros Lenguajes	419
fr	Francés	192
en gb	Inglés de Reino Unido	140

Tabla 4.3: Categorías de lenguaje y sus cantidades en la base de entrenamiento del modelo país

2. **Zona horaria**:

Para la zona horaria se consideran las que más apariciones tienen tanto en la clase de chilenos como en la clase de extranjeros. Realizando una unión, se obtiene un total de 34 categorías de zona horaria, 33 existentes y la última es la categoría 'otra'. En la siguiente Tabla 4.4 se pueden apreciar las zonas horarias que se incorporaron junto con su cantidad.

Ambos atributos se binarizan, quedando finalmente matrices binarias por categoría de cada atributo. Al unir las matrices de cada atributo se genera el input para entrenar los modelos necesarios. Por el lado del modelo país, queda una matriz con las *features* que provienen

Zona horaria	Zona horaria
santiago	mountain time (us & canada)
atlantic time (canada)	madrid
pacific time (us&canada)	arizona
quito	new delhi
brasilia	mexico city
caracas	casablanca
buenos aires	atlantic time (us & canada)
eastern time (us & canada)	rome
hawaii	pretoria
greenland	edinburgh
central time (us & canada)	alaska
america/santiago	mid-atlantic
athens	bogota
london	tokyo
amsterdam	west central africa
paris	

Tabla 4.4: Categorías de zona horaria en la base de entrenamiento del modelo país

del procesamiento de texto del conjunto de *tweets* y del location, además de *features* que vienen del lenguaje del usuario y zona horaria. En el caso del modelo región, como se dijo anteriormente, solo se utiliza aquella matriz obtenida del procesamiento de texto del conjunto de *tweets*.

4.3. Modelo país

Teniendo construida la matriz de *features* para utilizar como input en los diferentes algoritmos de *Machine Learning* que se aplican, se procede a entrenar los modelos y se escogerá aquel que reporte mejores resultados de desempeño. En el proceso de detección del mejor modelo, se analizan diferentes algoritmos y diferentes variaciones de incorporación de *features* para así observar la contribución de ciertas variables al modelo y elegir el mejor. Estas variaciones por cada algoritmo se resumen en la siguiente tabla:

Se consideran estas variaciones para analizar las diferencias entre utilizar sólo atributos de contenido (*tweets*), sólo atributos de contexto (lenguaje, *location* y zona horaria) o ambas. Se agrega también una componente de si cuando se utiliza atributo de contenido, este se preprocesa o no. Todo lo anterior se ve reflejado en las variaciones 1, 2, 3 y 4 y 5. En las variaciones 6, 7 y 8 se replican las variaciones 3, 4 y 5 pero eliminando el uso de la zona horaria ya que desde el 23 de Mayo 2018 este atributo no aparece en el JSON de la API REST de Twitter gratuitamente, por lo que se busca analizar si su uso en el modelo es suficientemente significativo para la clasificación. Para la comparación de los diferentes algoritmos y variaciones se utilizaron los parámetro que están por defecto para cada algoritmo en *Python*, los cuales se explicitan en la Tabla 4.6.

Variación	*tweets* con pre-procesamiento	*tweets* sin pre-procesamiento	location	lenguaje	zona horaria
1	Si	-	-	-	-
2	-	Si	-	-	-
3	Si	-	Si	Si	Si
4	-	Si	Si	Si	Si
5	-	-	Si	Si	Si
6	Si	-	Si	Si	-
7	-	Si	Si	Si	-
8	-	-	Si	Si	-

Tabla 4.5: Variaciones de *features* que se aplican para cada algoritmo a entrenar

	Parámetro	Valor	Descripción
MNB	alpha	1.0	Corrección de Laplace
	fit_prior	True	
	class_prior	None	
SGD	loss	hinge	Función de pérdida 'hinge' para entrenar SVM lineal.
	penalty	l2	Regularizador estándar de modelos SVM.
	alpha	0.0001	Constante que multiplica al regularizador
	l1_ratio	0.15	Parámetro de mezcla Elastic Net
	fit_intercept	True	Calcula el intercepto
	max_iter	5	Número máximo de pases en set entrenamiento
	tol	None	Critero para parar
	shuffle	True	Mezcla de datos entrenamiento después de cada paso.
	n_jobs	1	Número de CPUs que se usan para OVA [2]
	random_state	46	Semilla
	learning_rate	optimal	Tasa de aprendizaje
	power_t	0.5	Exponente para tasa de aprendizaje de escala inversa.
	n_iter	None	Nro de pasos sobre datos de entrenamiento (*epochs*)
RF	n_estimators	10	Número de árboles en el bosque
	criterion	gini	Función para medir calidad de división: Impureza gini
	max_features	auto	Nro. de *features* a considerar igual a sqrt(n_features)
	max_depth	None	Nodos se expanden hasta que hojas sean puras
	min_samples_split	2	Nro. mín. de muestras para dividir nodo interno
	min_samples_leaf	1	Nro. mín. de muestras requeridas para nodo hoja
	min_weight_fraction_leaf	0	
	max_leaf_nodes	None	Nodos hoja ilimitados
	min_impurity_decrease	0	Nodo se divide si disminución de impureza >= al valor.
	bootstrap	True	Usa muestras bootstrap en los árboles
	oob_score	False	No usa muestras *out-of-bag*
	n_jobs	1	Nro. de trabajos a correr paralelamente
	random_state	2375	Semilla
	verbose	0	Controla verbosidad del árbol
	warm_start	False	No reusa solución anterior para ajustar nuevo bosque
	class_weight	None	Peso 1 para todas las clases

Tabla 4.6: Parámetros por defecto de los diferentes algoritmos que se utilizan

Previo al entrenamiento de modelos, es importante recalcar que si bien se obtendrá Accuracy, Precision, Recall, F1-score y AUC como métricas de desempeño, aquellas que más se considerarán son el F1-Score de la clase 'Chile', AUC de la clase 'Chile' y el Accuracy general. Esto ya que el F1-Score captura el desempeño ponderado entre Precision y Recall de cada clase y 'Chile' es la clase de interés; el Accuracy general es aquel que nos indica la cantidad de ejemplos que están bien clasificados, tanto positivos como negativos, y como las clases están balanceadas es un buen indicador; y finalmente el AUC es aquel que es más útil para poder comparar entre modelos poniéndose en diferentes umbrales. En este caso, como las clases están balanceadas, el Accuracy nos permite sacar buenas conclusiones de clasificación sobre el modelo. Por lo tanto, el modelo que reporte un F1-Score de la clase Chile, AUC de la clase 'Chile' y Accuracy general más altos, entonces será el que se considerará como modelo país.

A continuación, se exponen los resultados de las 8 variaciones de los 3 algoritmos que se utilizan, para luego determinar cuál es el modelo que mejor realiza la labor de clasificación de usuarios chilenos *vs* usuarios extranjeros de *Twitter*. Cabe destacar que la base de datos construida para el modelo país se evalúa utilizando validación cruzada con 5 iteraciones, es decir, particiona aleatoriamente la base en 5 partes, entrenando con cuatro partes (80 %) y testeando en la parte restante (20 %).

4.3.1. Modelos entrenados y sus variaciones

Multinomial Naïve Bayes (MNB)

Para el caso de Naïve Bayes, se entrenan las 8 variaciones mencionadas en la Tabla 4.5 mediante el clasificador multinomial de este algoritmo. Esto, ya que tal como se menciona anteriormente, se cuenta con una matriz que considera valores tf-idf para cada palabra identificada en el proceso de tokenización del conjunto de los *tweets*, por lo que para poder procesar dichos valores es necesario utilizar Multinomial Naïve Bayes ya que Naïve Bayes sólo permite la utilización de un vector que indique presencia o ausencia de cada termino y no un valor como los que se obtienen con la matriz tf-idf. Se obtiene este clasificador desde la librería Sklearn de Python en el método 'naive_bayes'.

A continuación en la Tabla 4.7 se observan las métricas de desempeño obtenidas para cada variación tras el entrenamiento y testeo de los datos con validación cruzada de 5 iteraciones. En ella, se puede identificar *precision*, *recall* y *F1-score* de cada clase, junto con *accuracy* del modelo general y AUC en base a la clase positiva, la cual es 'Chile'.

En la Tabla 4.8 se aprecia que la desviación estándar entre los resultados de las cinco iteraciones de entrenamiento con MNB, si bien no es tan alta son los valores que dan más alto en comparación a los otros algoritmos, como se verá más adelante. Por ende, el modelo es medianamente independiente con respecto a la partición de datos que se realiza para el entrenamiento según los datos con los que se cuenta.

Variación	Clases						Acc. (%)	AUC (%)
	Chile			Otro				
	Prec. (%)	Rec. (%)	F1 (%)	Prec. (%)	Rec. (%)	F1 (%)		
1	85,04	93,28	88,97	92,51	83,48	87,76	88,40	91,89
2	85,02	92,63	88,66	91,85	83,57	87,51	88,12	91,80
3	86,75	98,45	92,23	98,20	84,87	91,05	91,68	98,39
4	87,17	98,63	92,55	98,42	85,39	91,44	92,03	98,54
5	**93,85**	**99,29**	**96,49**	**99,24**	**93,45**	**96,26**	**96,38**	**99,74**
6	86,31	98,18	91,86	97,87	84,32	90,59	91,27	97,87
7	86,43	98,41	92,03	98,14	84,44	90,78	91,45	98,05
8	91,72	99,14	95,28	99,05	90,99	94,85	95,08	99,67

Tabla 4.7: Métricas de desempeño de las ocho variaciones entrenadas con Multinomial Naïve Bayes

	Valor promedio (%)	Mín. (%)	Máx. (%)	Desv. Est. (%)
Accuracy	96,38	93,19	98,25	1,70
F1-Score	96,37	93,16	98,25	1,71
Precision	96,59	93,87	98,27	1,47
Recall	96,37	93,17	98,25	1,71
AUC	99,74	99,40	99,94	0,18

Tabla 4.8: Estadísticas descriptivas de las métricas de desempeño de la Variación 5 - MNB en las 5 iteraciones de validación cruzada

Stochastic Gradient Descent (SGD)

Para el caso de Support Vector Machine, se utiliza el clasificador Stochastic Gradient Descent el cual es una versión de SVM que considera los gradientes en puntos aleatorios para minimizar la distancia de cada instancia a los hiperplanos y así lograr el objetivo de clasificar según distancia mínima; el cual se utiliza por su buen comportamiento frente a una alta cantidad de datos y frente a la utilización de la matriz tf-idf ya que funciona bien al considerar datos dispersos. Se utiliza *SGDClassifier* de la librería de Sklearn de Python en el método 'linear_model'

A continuación se observan las métricas de desempeño obtenidas para cada variación tras el entrenamiento y testeo de los datos con validación cruzada de cinco iteraciones. Se identifican las mismas métricas que el algoritmo anterior.

Variación	Clases						Acc. (%)	AUC (%)
	Chile			Otro				
	Prec. (%)	Rec. (%)	F1 (%)	Prec. (%)	Rec. (%)	F1 (%)		
1	90,53	96,27	93,31	95,99	89,86	92,82	93,08	98,63
2	89,97	96,01	92,89	95,69	89,22	92,34	92,63	98,48
3	98,33	98,85	98,59	98,83	98,31	98,57	98,58	99,89
4	98,48	98,92	98,70	98,91	98,46	98,68	98,69	99,90
5	97,81	98,45	98,13	98,43	97,79	98,11	98,12	99,79
6	**98,91**	**99,00**	**98,95**	**98,99**	**98,90**	**98,95**	**98,95**	**99,89**
7	98,67	98,97	98,82	98,96	98,66	98,81	98,81	99,89
8	97,03	98,24	97,63	98,21	96,98	97,59	97,61	99,76

Tabla 4.9: Métricas de desempeño de las ocho variaciones entrenadas con Stochastic Gradient Descent

En la Tabla 4.10, se aprecian las estadísticas relevantes con respecto a los diferentes resultados de las métricas de desempeño en las cinco iteraciones de validación cruzada. Se observa una baja desviación estándar entre estos valores en todas las métricas, por lo que se puede decir que este modelo (SGD - Variación 6) se comporta bastante bien independiente de la partición que se realice para entrenar y luego testear.

	Valor promedio (%)	Mín. (%)	Máx. (%)	Desv. Est. (%)
Accuracy	98,95	97,76	99,62	0,63
F1-Score	98,95	97,76	99,62	0,63
Precision	98,96	97,77	99,62	0,62
Recall	98,95	97,75	99,62	0,63
AUC	99,89	99,71	99,99	0,1

Tabla 4.10: Estadísticas descriptivas de las métricas de desempeño de la Variación 6 - SGD en las 5 iteraciones de validación cruzada

Random Forest (RF)

Se entrenan las ocho variaciones con Random Forest mediante el algoritmo *RandomForestClassifier* de la librería Sklearn de Python en el método 'ensemble'.

A continuación, se observan las métricas de desempeño que se obtienen tras el entrenamiento y testeo de cada una de las ocho variaciones utilizando validación cruzada con cinco iteraciones. Se identifican las mismas métricas que los algoritmos anteriores.

Variación	Clases						Acc. (%)	AUC (%)
	Chile			Otro				
	Prec. (%)	Rec. (%)	F1 (%)	Prec. (%)	Rec. (%)	F1 (%)		
1	86,19	94,93	90,35	94,32	84,69	89,24	89,30	95,63
2	85,43	95,57	90,21	94,94	83,59	88,90	89,60	95,41
3	87,09	96,01	91,34	95,52	85,68	90,33	90,86	96,58
4	85,72	95,81	90,49	95,22	83,49	89,22	89,90	95,96
5	**98,01**	**97,65**	**97,83**	**97,64**	**98,00**	**97,82**	**97,82**	**99,64**
6	87,42	95,46	91,26	94,97	86,17	90,35	90,83	96,49
7	85,99	95,74	90,61	95,16	84,31	89,40	90,04	95,92
8	97,42	97,30	97,36	97,29	97,40	97,35	97,35	99,69

Tabla 4.11: Métricas de desempeño de las ocho variaciones entrenadas con Random Forest

A continuación, en la Tabla 4.12 se observan algunas estadísticas descriptivas sobre las métricas de desempeño evaluadas en las 5 iteraciones de validación cruzada en el entrenamiento y testeo del modelo con Random Forest. A partir de éstas, se puede observar que su comportamiento es bastante homogéneo independiente de la muestra de entrenamiento que se considere. Esto ya que la desviación estándar de las métricas en las 5 iteraciones es bastante baja (0,17 % para el AUC y 0,13 % para las demás métricas) por lo que los valores se mantienen muy constantes en cada iteración.

	Valor promedio (%)	Mín. (%)	Máx. (%)	Desv. Est. (%)
Accuracy	97,82	97,64	97,98	0,13
F1-Score	97,82	97,64	97,98	0,13
Precision	97,84	97,64	97,99	0,13
Recall	97,82	97,64	97,99	0,13
AUC	99,64	99,33	99,99	0,17

Tabla 4.12: Estadísticas descriptivas de las métricas de desempeño de la Variación 5 - RF en las 5 iteraciones de validación cruzada

Variación	Publicaciones con preprocesamiento	Publicaciones sin preprocesamiento
1	Si	-
2	-	Si

Tabla 4.13: Variaciones de atributo de contenido a aplicar en cada algoritmo a entrenar

4.4. Modelo región

Para el diseño de este modelo sólo se considera la variable de contenido, es decir, publicaciones de los usuarios. Es por esto que sólo se analizan dos variaciones para evaluar mejoras en desempeño en cada algoritmo. Éstas son con y sin preprocesamiento de texto, tal como se muestra en la Tabla 4.13:

Al igual que el caso del modelo país, para encontrar el modelo que tenga mejor desempeño de clasificación de las diferentes clases se prueban tres algoritmos: Multinomial Naïve Bayes, dado que es un modelo multiclase y que además considera matriz tf-idf como input; Stochastic Gradient Descent, también se comporta bien con problemas multiclase; y Random Forest.

Para entrenar los tres algoritmos recién mencionados también se utiliza Python, por lo que los parámetros por defecto con los que se prueban los modelos son los mismos que se mencionan anteriormente en la Tabla 4.6.

Un punto relevante a considerar en el modelo región es la cantidad de clases que este posee. En este caso se tienen quince clases por las quince regiones administrativas de Chile, y por lo tanto no existe una clase positiva ya que el objetivo es clasificar correctamente todas las clases. Teniendo en cuenta que la base de entrenamiento a utilizar está medianamente desbalanceada, ya que naturalmente se identificaron muchos más usuarios residentes de la Región Metropolitana que de otras regiones, por ejemplo, y poder capturar esta diferencia parece importante para que el aprendizaje automático de cada región sea mucho más completo. Es por esto que es fundamental tener en consideración además de Accuracy, el desempeño por región. Esto ya que si un modelo tiene un alto Accuracy pero principalmente por clasificar muy bien sólo las regiones con más registros, por ejemplo, quizás no necesariamente sea mejor que uno clasifique todas las regiones con un *F1-Score* similar pero que entregue Accuracy levemente menor. Como el siguiente problema de modelamiento es multiclase, entonces no se contará con un solo AUC si no que con una métrica AUC por cada clase.

Para el entrenamiento del modelo región, se utiliza la técnica *One-Vs-Rest* explicada en el Capítulo 2 de manera que se entrene un clasificador por cada clase para lograr una mejor identificación de cada clase con respecto a las otras considerando la cantidad de clases con la que se cuenta. Para llevar a cabo lo anterior, se utiliza *OneVsRestClassifier* del método *multiclass* de la librería Sklearn de Python. Cabe destacar que *SGDClassifier* incorpora por defecto la opción de One-Vs-Rest para los modelos multiclase, por lo que en dicho caso no es necesario agregar el clasificador de *One-Vs-Rest*.

A continuación, se exponen los resultados de las dos variaciones de los tres algoritmos que se utilizan, para luego determinar cuál es el modelo que mejor realiza la labor de clasificación en las quince regiones de Chile. En esta oportunidad también se evalúa el entrenamiento y

testeo de los modelos mediante validación cruzada de cinco iteraciones.

4.4.1. Modelos entrenados y sus variaciones

Multinomial Naïve Bayes (MNB)

Se entrenan las dos variaciones con el clasificador *MultinomialNB* de la librería Sklearn de Python capturando las métricas Precision, Recall y F1-Score por cada clase, y Precision Macro, Recall Macro, F1-Score Macro y Accuracy del modelo en general.

En la Tabla 4.14 se observan los resultados por cada variación de todas las métricas recién mencionadas.

Clases	Variación					
	1			2		
	Prec. (%)	Rec. (%)	F1-Score (%)	Prec (%)	Rec (%)	F1-Score (%)
1	87,70	33,48	48,46	88,66	32,66	47,74
2	78,06	23,59	36,23	79,00	23,10	35,75
3	82,70	17,21	28,49	86,67	16,85	28,21
4	68,35	13,23	22,17	70,08	12,40	21,07
5	49,77	12,54	20,03	54,21	13,53	21,65
6	55,98	15,93	24,80	64,95	14,63	23,88
7	58,33	26,75	36,68	60,97	25,94	36,39
8	19,05	39,75	25,76	20,45	38,22	26,65
9	72,82	26,46	38,81	77,93	24,47	37,42
10	53,33	17,07	25,86	56,36	16,63	25,68
11	42,43	49,78	45,82	41,51	49,93	45,33
12	75,06	22,32	34,41	80,39	21,37	33,76
13	14,58	80,99	24,72	14,24	83,76	24,34
14	71,71	21,81	33,45	72,07	21,44	33,05
15	59,39	10,07	17,22	64,92	9,18	16,08
Prec. Macro	59,63 %			62,38 %		
Rec. Macro	27,40 %			26,94 %		
F1 Macro	**30,85 %**			30,45 %		
Accuracy	**29,23 %**			28,90 %		

Tabla 4.14: Métricas de desempeño de las dos variaciones entrenadas con Multinomial Naïve Bayes

Con respecto al desempeño de la Variación 1 - MNB, en la Tabla 4.15 se observa el comportamiento de las métricas de desempeño del modelo en las cinco iteraciones de validación cruzada. Realizando un intervalo de confianza al 95 % se puede apreciar que las métricas se mueven a lo más 2 % aproximadamente. Por lo tanto, es un modelo que no varía considerablemente según la partición que se realice en los datos para el entrenamiento y testeo. Esto

indica que es un modelo más o menos estable y que al considerar una cierta partición al momento de entrenar, los resultados al validar el modelo son bastante similares a que si se hubiese realizado otra partición en los mismos datos.

	Valor promedio (%)	Mín (%)	Máx (%)	Int. Confianza (%)
Accuracy	29,23	28,07	30,35	+/- 1,45
Precision Macro	59,63	57,90	60,65	+/- 2,01
Precision Micro	29,23	28,07	30,35	+/- 1,45
Recall Macro	27,40	26,22	28,51	+/- 1,46
Recall Micro	29,23	28,07	30,35	+/- 1,45
F1-Score Macro	30,85	29,82	32,10	+/- 1,47
F1-Score Micro	29,23	28,07	30,35	+/- 1,45

Tabla 4.15: Estadísticas de las métricas de desempeño de la Variación 1 - MNB en las cinco iteraciones de validación cruzada

Stochastic Gradient Descent (SGD)

En cuanto a este algoritmo, también se entrenan los dos Variaciones planteadas anteriormente las cuales arrojan los siguientes resultados de la Tabla 4.16:

En la Tabla 4.17 se observan algunas estadísticas con respecto a los valores obtenidos para las métricas de desempeño en las cinco iteraciones de validación cruzada.

Con respecto a dichas estadísticas, se aprecia un intervalo de confianza más alto que lo que reportaba validación cruzada con Multinomial Naïve Bayes. Si se observa *F1-Score Macro* este tiene un intervalo de +/- 2,38 lo que lleva a que, dependiendo de la partición que se realice para el entrenamiento, esta métrica podría ir entre los valores 40,16 % y 44,92 %, intervalo el cual sigue siendo mejor que el reportado para la Variación 1 - MNB.

Random Forest (RF)

Finalmente, al entrenar este último algoritmo con las dos variaciones planteadas se obtienen los resultados que se muestran en la Tabla 4.18.

A continuación se presentan las estadísticas con respecto a la Variación 1 - RF de los diferentes valores obtenidos en las métricas en las cinco iteraciones de validación cruzada.

El intervalo de confianza que se observa en todas las métricas de la Tabla 4.19 es el menor que se ha reportado dentro de los tres algoritmos que se presentan. Esto indica que Random Forest es el algoritmo más estable con respecto a la independencia de la partición de set de entrenamiento y testeo. Lo mismo ocurre en el caso del modelo país.

Clases	Variación					
	1			2		
	Prec. (%)	Rec. (%)	F1-Score (%)	Prec (%)	Rec (%)	F1-Score (%)
1	58,86	49,07	53,52	62,87	48,10	54,50
2	53,72	40,33	46,07	51,71	41,17	45,84
3	44,47	38,23	41,11	47,58	38,30	42,44
4	47,91	38,23	42,53	46,96	38,72	42,44
5	25,35	33,00	28.67	27,56	35,28	30,95
6	40,39	36,21	38,18	35,96	37,73	36,82
7	53,45	41,00	46,40	53,94	41,22	46,73
8	42,67	34,03	37,86	24,82	47,75	32,67
9	31,28	49,08	38,21	31,38	48,64	38,15
10	46,29	39,51	42,64	52,48	39,00	44,74
11	63,95	53,48	58,25	65,23	53,84	58,99
12	49,65	46,33	47,93	47,62	46,92	47,26
13	23,79	40,68	30,02	25,21	22,14	23,57
14	46,30	38,50	42,04	46,40	38,28	41,95
15	44,48	36,71	40,23	44,28	36,64	40,10
Prec. Macro	47,54			47,90		
Rec. Macro	40,96			40,91		
F1 Macro	**42,54**			42,35		
Accuracy	**40,73**			40,21		

Tabla 4.16: Métricas de desempeño de las dos variaciones entrenadas con Stochastic Gradient Descent

	Valor promedio (%)	Mín. (%)	Máx. (%)	Int.Confianza (%)
Accuracy	40,73	38,15	42,30	+/- 3,36
Precision Macro	47,54	46,08	48,86	+/- 2,09
Precision Micro	40,73	38,15	42,30	+/- 3,36
Recall Macro	40,96	39,55	41,95	+/- 1,88
Recall Micro	40,73	38,15	42,30	+/- 3,36
F1-Score Macro	42,54	40,73	43,88	+/- 2,38
F1-Score Micro	40,73	38,15	42,30	+/- 3,36

Tabla 4.17: Estadísticas de las métricas de desempeño de la Variación 1 - SGD en las cinco iteraciones de validación cruzada

Clases	Variación					
	1			2		
	Prec. (%)	Rec. (%)	F1-Score (%)	Prec (%)	Rec (%)	F1-Score (%)
1	61,88	39,22	48,01	59,74	34,75	43,94
2	50,80	28,75	36,72	51,40	26,87	35,29
3	46,46	26,93	34,09	44,51	26,28	33,05
4	42,80	23,40	30,26	39,38	22,84	28,91
5	35,23	15,92	21,93	32,84	14,11	19,74
6	40,87	28,67	33,70	41,69	28,17	33,62
7	45,35	32,99	38,20	43,99	33,87	38,27
8	38,06	22,95	28,64	35,77	22,58	27,69
9	42,45	37,73	39,95	41,19	33,75	37,10
10	37,79	32,01	34,66	35,12	29,95	32,33
11	50,62	50,07	50,35	48,69	48,55	48,62
12	38,19	40,01	39,08	36,53	40,60	38,46
13	19,97	57,11	29,60	20,09	56,05	29,58
14	30,11	38,35	33,74	28,47	35,91	31,76
15	23,49	34,20	27,85	20,91	33,46	25,73
Prec. Macro	40,34			36,27		
Rec. Macro	33,89			31,26		
F1 Macro	35,10			33,58		
Accuracy	34,26			32,92		

Tabla 4.18: Métricas de desempeño de las dos variaciones entrenadas con Random Forest

	Valor promedio (%)	Mín. (%)	Máx. (%)	Int. Confianza (%)
Accuracy	34,26	33,43	35,12	+/- 1,12
Precision Macro	40,34	39,52	41,21	+/- 1,33
Precision Micro	34,26	33,43	35,12	+/- 1,12
Recall Macro	33,89	32,99	34,70	+/- 1,15
Recall Micro	34,26	33,43	35,12	+/- 1,12
F1-Score Macro	35,10	34,19	35,68	+/- 1,14
F1-Score Micro	34,26	33,43	35,12	+/- 1,12

Tabla 4.19: Estadísticas de las métricas de desempeño de la Variación 1 - RF en las cinco iteraciones de validación cruzada

Capítulo 5

Análisis de resultados

En la sección 5.1 y 5.2 del presente capítulo, se analizan los resultados obtenidos del entrenamiento y testeo de modelos, comparándolos entre sí para así determinar aquel que mejor desempeño obtiene según el problema planteado.

Luego, en la sección 5.3 se detalla la evaluación de desempeño de la heurística que se utiliza hoy en día para clasificar cuentas de usuarios de *Twitter* que viven en Chile para luego en la sección 5.4 realizar una comparación de desempeño entre el modelo país elegido con la heurística, para así validar si efectivamente logra capturar más cuentas de usuarios chilenos con un nivel de precisión igual o mejor.

5.1. Evaluación de desempeño de modelo país

5.1.1. Análisis MNB

Se puede apreciar de la Tabla 4.7 que si se utiliza sólo la variable de *tweets* (Variación 1 y 2), el modelo que tiene mejores resultados es aquel en que se preprocesa el texto de los *tweets* (Variación 1) *versus* que no se preprocesen (Variación 2). Pero, cuando se agregan otras variables del usuario aparte de sus publicaciones (Variación 3, 4, 6 y 7), los resultados son mejores que si sólo se utiliza contenido. Pero esta vez los modelos sin *tweets* preprocesados (Variación 4 y 7) entregan mejores resultados. Comparando la variación 3 con la 4, el AUC sólo mejora 0,15 % en la variación 4, y comparando la variación 6 con la 7 la mejora es de 0,18 % en la variación 7 en relación a la misma métrica.

Por otro lado, la mejora de los modelos que consideran publicaciones del usuario junto con variables de contexto se puede apreciar comparando la Variación 1 con la 3 y la 6. La 1 considera sólo contenido preprocesado mientras que la 3 y la 6 contenido preprocesado y contexto, y la mejora de éstas últimas con respecto a la variación 1 es de 6,5 % y 5,98 %, respectivamente. La diferencia entre la Variación 3 y 6 es que la primera considera la zona horaria como una de las variables de contexto mientras que la última no. Se hace esta dife-

rencia ya que, dado que desde el 23 de Mayo 2018 la zona horaria no aparece en el JSON de la API REST que se puede extraer gratuitamente, se busca evaluar su significancia marginal y si sería relevante pagar por esta variable.

Se realiza el mismo análisis anterior, pero para el caso del contenido sin preprocesamiento. Las variaciones a comparar son la 2 (sólo contenido sin preprocesamiento) con la variación 4 y 7 (variables de contexto), en las cuales la 4 considera zona horaria y la 7 no. Al igual que en el caso anterior, se obtienen mejores resultados en las Variaciones que incluyen variables de contexto. Y comparando ambas Variaciones (4 y 7), la número 4 mejora en un 0,49 % el AUC y 0,76 % en el Accuracy general, aunque disminuye 0,42 % el *F1-Score* de la clase 'Chile'.

En el caso de las Variaciones 5 y 8, éstas sólo consideran en el modelo variables de contexto. La diferencia entre ambas es que la número 5 considera la zona horaria mientras que la 8 no. En cuanto a desempeño, la Variación 5 es mejor que la número 8, por lo que incluir la zona horaria genera una mejora de 0,7 % en el AUC y 1,21 % en el *F1-Score* de la clase 'Chile'.

Para el algoritmo Naïve Bayes Multinomial la utilización de variables de contexto genera una mejora en los resultados. En los casos en que se considera la zona horaria, los mejores resultados que se obtienen son en el modelo que utiliza sólo variables de contexto y sin contenido (Variación 5). Y para los casos en que no se considera la zona horaria, también da mejores resultados aquel modelo que sólo utiliza variables de contexto (Variación 8).

Finalmente, para MNB la Variación 5 es aquella que tiene mejores resultados ya que contiene los valores más altos en toda las métricas señaladas. Cabe recordar, que la variación número 5 es aquel modelo que utiliza sólo las variables *location*, lenguaje y zona horaria para crear las *features*. Es decir, no se utilizan los *tweets* publicados por los usuarios para la clasificación. Y la Variación 8 es el modelo que mejor resultados tiene cuando no se utiliza la zona horaria.

5.1.2. Análisis SGD

Se puede apreciar en la Tabla 4.9 que al igual que para el caso de MNB, si sólo se utiliza la variable de contenido (Variación 1 y 2), el modelo que arroja mejores resultados es aquel de la Variación 1, es decir, clasifica mejor el que preprocesa los *tweets*. Dicha Variación mejora un 0,15 % la métrica AUC, y esta diferencia se debe principalmente al *F1-Score* de la clase 'Otro' ya que este cambia de 92,34 % a 92,82 %, es decir mejora un 0,48 % al preprocesar las publicaciones de los usuarios. Aunque el *F1-Score* de la clase 'Chile' igual mejora en un 0,42 %.

Cuando se incorporan variables de contexto (Variación 3, 4, 6 y 7), ya sea con o sin preprocesamiento de los *tweets*, los resultados mejoran aproximadamente 1,3 % en el AUC en comparación a cuando sólo se utiliza la variable de contenido. Las cuatro variaciones recién mencionadas arrojan una métrica AUC bastante similar (99,89 - 99,90 %) pero el que tiene mejores resultados de *F1-Score* de la clase 'Chile' se diferencia del peor principalmente por la precisión de la clase 'Chile' que aumenta un 0,58 % y por el *recall* de la clase 'Otro' que aumenta un 0,59 % (Variación 3 *versus* Variación 6). Cabe destacar que la Variación 3 es

aquella que considera *tweets* preprocesados con las tres variables de contexto, mientras que la Variación 6 es lo mismo sólo que no considera la zona horaria. Por ende, en este caso, al utilizar la zona horaria los resultados de clasificación empeoran.

Si se compara la Variación 3 con la 4, que se diferencian por el preprocesamiento de los *tweets*, la Variación que mejor resultados entrega es aquella que no preprocesa la variable de contenido; mientras que si se compara la Variación 6 y 7, las cuales no consideran zona horaria, la Variación que mejor resultados entrega es con preprocesamiento. De todas formas, tal como se dijo anteriormente, la Variación 6 es la que tiene mejor desempeño en la clasificación.

Por último, en el caso de las Variaciones 5 y 8, las cuales sólo consideran variables de contexto, aquella que tiene mejor desempeño es la que considera la zona horaria como variable de contexto (Variación 5) mejorando sólo un 0,03 % en el AUC, pero un 0,5 % en el *F1-Score* de 'Chile'.

Aún así, y a diferencia de los resultados obtenidos con MNB, en este caso la Variación que obtiene un mejor desempeño de clasificación es la número 6 la cual considera variable de contenido con preprocesamiento y variables de contexto sin zona horaria.

5.1.3. Análisis RF

En la Tabla 4.11 se observa que, al igual que MNB y SGD, la Variación 1 arroja mejores resultados que la Variación 2, considerando que ambos modelos utilizan sólo la variable de contenido y se diferencian uno del otro por el preprocesamiento de dicha variable. Esta mejora de la Variación 1 es de 0,22 % en el valor AUC el cual está dado principalmente por el aumento de un 0,34 % en el *F1-Score* de la clase 'Otro'.

Si se agregan a los modelos las variables de contenido (Variación 3, 4, 6 y 7), todas mejoran con respecto a la 1 y 2, por lo que agregar variables de contexto aparte de la variable de contenido genera un aporte en la tarea de clasificación de usuarios chilenos. Además, se observa que en las cuatro variaciones antes mencionadas (3, 4 6 y 7) aquellas que contienen la variable de contenido preprocesada (Variación 3 y 6) tienen mejores resultados que las otras (Variación 4 y 7), por lo que cuando se agregan las variables de contenido, es mejor preprocesar las publicaciones del usuario para mejorar las métricas de desempeño del modelo.

Por otro lado, para los casos en que se preprocesan las publicaciones y se agregan variables de contenido (Variación 3 y 6), agregar la zona horaria dentro de las variables de contenido (Variación 3) mejora levemente el *F1-Score* de la clase 'Chile', Accuracy y AUC.

Al comparar la Variación 6 con la 7, se observa que preprocesar las publicaciones de los usuarios genera mejores resultados cuando no se utiliza la zona horaria como variable de contexto. Esta mejora es de un 0,57 % en el AUC y de 0,65 % en el *F1-Score* de la clase 'Chile'.

Continuando con el análisis, si sólo se consideran las variables de contexto al momento de modelar el problema, ocurre lo mismo que con los algoritmos anteriores en cuanto a que

agregar la zona horaria aporta en la tarea de clasificación. Esto ya que la Variación 5 presenta un AUC levemente mayor pero un *F1-Score* de 'Chile' 0,47 % mayor, por sobre la Variación 8.

En resumen, si se utiliza la variable de contenido, incorporar también variables de contexto definitivamente lleva a obtener resultados mejores cuando se modela con Random Forest. Además, al utilizar ambos tipos de información de usuario, se observa que existe una leve mejora en las métricas de desempeño al preprocesar los *tweets* y *retweets*. Pero, los mejores resultados para el caso de Random Forest son obtenidos cuando no se utiliza la variable de contenido, es decir, cuando no se consideran los *tweets* y *retweets* que realiza el usuario dentro del modelo, ya que la Variación N° 5 es con la que se obtienen mejores resultados tanto en todas las métricas que se evalúan.

5.1.4. Análisis General

Tras comparar los modelos entrenados mediante 3 algoritmos diferentes y con 8 variaciones (Tablas 4.7, 4.9 y 4.11), se obtiene finalmente que el modelo que mejor realiza la tarea de clasificación de usuarios chilenos *versus* usuarios extranjeros de *Twitter* es la Variación N° 6 de Stochastic Gradient Descent (SGD), es decir, aquel modelo que utiliza los *tweets* y *retweets* publicados por el usuario preprocesados, junto con las variables de contenido: *location* y *lenguaje*.

Si bien inicialmente se pensaba que la zona horaria sería un aporte significativo en la clasificación de los usuarios chilenos de *Twitter*, asombrosamente el modelo que arroja mejores métricas de desempeño con validación cruzada no incorpora dicha variable. Este resultado es bastante favorable ya que no se requiere del dato de la zona horaria de cada usuario para obtener los mejores resultados de clasificación, por ende, que *Twitter* haya bloqueado el acceso gratuito a esta variable a través de la API REST de dicha plataforma no afecta en el modelo encontrado. De todas formas, aunque utilizar dicha variable hubiese mejorado las métricas de desempeño del modelo, tal como se puede observar en las tablas de resultados los valores de las métricas no difieren tanto.

Si se considera el segundo mejor modelo, este sería también con el algoritmo Stochastic Gradient Descent pero en su Variación número 7, es decir, lo mismo que la Variación 6 sólo que las publicaciones de los usuarios no se preprocesan. Este modelo tiene el mismo valor AUC (99,89 %) pero disminuye un 0,13 % en el *F1-Score* de 'Chile', un 0,14 % en el *F1-Score* de 'Otro' y un 0,14 % en Accuracy. Esto en términos de la cantidad de clasificados afecta en que se en la Variación 7 se clasifican mal 8 usuarios más que en la Variación 6, y además la cantidad de chilenos clasificados correctamente disminuye en 16. Tomando en cuenta que el preprocesamiento no es un proceso demoroso ya que no toma más de 1 minuto en ejecutar las diferentes etapas del preprocesamiento, entonces se opta por la Variación 6 aunque en términos porcentuales con respecto al total de la base la mejora es bastante poca.

Por otro lado, cabe recordar que la presente investigación se realiza exclusivamente con usuarios activos de *Twitter*, es decir, usuarios que tienen al menos una publicación (*tweet* o *retweet*) en el periodo de 1 año. En este caso, como se comenzaron a extraer datos para

construir la base de entrenamiento el 13 de mayo de 2018, se consideró el periodo entre el 12 de mayo 2017 y 12 de mayo 2018. Considerando sólo a los usuarios activos se podría evaluar el impacto de utilizar la variable de contenido al momento de clasificar. Pero, teniendo en cuenta que se consideran sólo usuarios activos en la base de entrenamiento, si se observa la Variación 8 de Stochastic Gradient Descent, que es uno de los modelos que también clasifica bastante bien los casos chilenos versus no chilenos, se puede concluir que con sólo las variables de lenguaje y campo *location* se obtienen resultados bastante parecidos a que si se utilizan las publicaciones del usuario. En contraste con la Variación 6 - SGD, la Variación 8 - SGD disminuye un 0,13 % en AUC, un 1,32 % en *F1-Score* de 'Chile', un 1,36 % en *F1-Score* de 'Otro' y un 1,34 % en Accuracy. Esto, permite cuestionarse el real aporte que generan las publicaciones en la búsqueda de usuarios chilenos en el Universo de *Twitter*. Si bien estos resultados no son extrapolables a todo el Universo de *Twitter* ya que se estaría asumiendo que los usuarios en cuestión que realizaron al menos una publicación en el año tienen el mismo comportamiento que aquellos que no publicaron en dicho año, en relación a las variables del lenguaje de la interfaz y lo que ingresan en su campo location; sería interesante analizarlo para evaluar el real aporte de incorporar las publicaciones para identificar a los usuarios chilenos.

Dado lo anterior, entonces el mejor modelo para realizar la clasificación de usuarios chilenos versus usuarios extranjeros en *Twitter* es la Variación 6 - SGD, el cual llamaremos modelo país.

Si bien el modelo mencionado anteriormente es aquel que arroja mejores resultados, los demás modelos con sus variaciones en general también arrojan buenos resultados y esto genera el cuestionamiento de la existencia de sobreajuste. Dado que se están considerando los *tweets* de los usuarios sin restricción de lenguaje, se podría pensar que el modelo con sólo discriminar entre español versus los demás lenguajes podría arrojar buenos resultados y de esta forma el aprendizaje automático no haría mucha diferencia. Es por esto, que se realiza un análisis descriptivo con respecto al lenguaje de la base de entrenamiento del modelo país y se observa que de los 6.301 chilenos de la base, 4.426 tienen como idioma prevalente el español [1]. Por otro lado, de los 6.462 extranjeros de la base, 1.101 tienen como idioma prevalente el español. Por ende, 41,89 % de la base tiene la mayoría de sus *tweets* en español, esto indicaría que si el modelo clasificara a los chilenos sólo según el lenguaje se tendría como desempeño un Accuracy de 0.752, precision 'Chile' de 0.673 y Recall 'Chile' de 0.794. Estas métricas si bien son buenas, no son mejores que el modelo encontrado, por lo que se rechaza la hipótesis de que el lenguaje esté provocando sobreajuste en la base.

Para validar los resultados obtenidos en el Capítulo anterior con respecto al modelo país, se construye una base de validación formada por 8.126 usuarios de los cuales 510 son chilenos, es decir, el 6 % aproximadamente de la base. Se construye con una baja cantidad de usuarios chilenos ya que estos son bastante pocos en relación a todo el Universo de *Twitter* por lo que esta base se acerca un poco más a la realidad que el 50/50 como fue entrenado el modelo. Además, es importante recalcar que para armar la base de validación sólo se utilizan usuarios geolocalizados, tanto chilenos como extranjeros.

[1] El idioma prevalente se determina según el idioma más frecuente en que cada usuario emitió sus *tweets* y *retweets*

A continuación en la Tabla 5.1 se muestran los resultados obtenidos tras utilizar el modelo país entrenado mediante validación cruzada en el capítulo anterior (Variación 6 - SGD) , para predecir qué usuarios de la base de validación son chilenos.

Clases	Modelo			
	Prec. (%)	Rec. (%)	F1-Score (%)	Support
Chile	94,90	83,92	89,07	510
Otro	98,93	99,70	99,31	7616
AUC	99,03 %			8126
F1-Score	98,67 %			
Accuracy	98,71 %			

Tabla 5.1: Validación de resultados modelo país SGD

Estos resultados son bastante buenos y dentro del rango de lo esperado con respecto al intervalo de confianza planteado en la Tabla 4.10 con respecto al modelo país. El Accuracy en este caso no aporta mucha información ya que gran parte de la base son usuarios extranjeros por lo que si estos están bien clasificados aumenta el *Accuracy* y en realidad lo que importa es la clasificación que se realiza con respecto a 'Chile'. Es por esto que a continuación se presenta la matriz de confusión de la clasificación de la base de validación en la Tabla 5.2.

		Predicha	
		Otro	Chile
Real	Otro	7593	23
	Chile	82	428

Tabla 5.2: Matriz de confusión modelo país (Variación 6 - SGD)

5.2. Evaluación de desempeño heurística actual de clasificación de usuarios chilenos

5.2.1. Descripción de heurística de clasificación

En el año 2016, se diseñó e implementó una heurística [12] para establecer una base de usuarios que fueran chilenos a partir de ciertas palabras que aparezcan en el campo *location* como 'Chile', nombres de las regiones (Metropolitana, Bío Bío, etc), nombres de comunas que hasta el CENSO 2012 contaran con más de 30.000 habitantes y variantes como 'shile', 'stgo', entre otras. Además de incorporar a aquellos que tuvieran ubicación geo referenciada y que dicho punto de coordenadas perteneciera a Chile.

Mediante dicha heurística se extrajo una base de usuarios chilenos que actualmente tiene 1.638.317 usuarios. Pero, se identifica que algunos de dichos usuarios no son chilenos ya que el algoritmo identifica ciertas palabras que pertenecen a lugares de Chile que en realidad en su

totalidad se refieren a un lugar de otra parte del mundo. Un ejemplo de esto es 'Santiago de Compostela, España', ya que contiene la palabra 'Santiago' y clasifica a esos usuarios como chilenos asumiendo que se refería a Santiago de Chile cuando en realidad es Santiago de Compostela de España. Esto provoca que la base de chilenos contenga usuarios extranjeros y que además sólo se puedan identificar a los usuarios chilenos que ingresen el campo *location*. Es por dichas razones que se diseña un modelo de *Machine Learning* que permita identificar a los usuarios chilenos de *Twitter* que considere otros atributos aparte del campo *location*, para así contar con una base de usuarios chilenos de *Twitter* más amplia y que además sea mediante aprendizaje automático.

Para analizar cuantitativamente si el modelo de *machine learning* que se diseña tiene o no mejores resultados que la heurística, es necesario primero evaluar el desempeño de la heurística tanto a nivel de precisión como de *recall*.

En cuanto a la precisión, aunque no es tarea fácil dada la cantidad de los datos, basta con analizar cuántos de los usuarios que la heurística clasificó como chilenos son efectivamente chilenos. Pero, en cuanto al *recall*, no es posible determinar cuántos usuarios chilenos no fueron clasificados como chilenos ya que no se tiene etiquetado todo el universo de *Twitter*, de hecho ese es el problema que tanto la heurística como el modelo de *Machine Learning* busca solucionar. Es por esto que se dividió el proceso de evaluación en dos etapas:

- Etapa 1: Evaluación de la precisión de los datos recuperados por la heurística en todo el universo de *Twitter*.
- Etapa 2: Evaluación de precisión y *recall* comparativo entre la heurística y el modelo país según los usuarios que cada uno recopile, a partir de una base de datos etiquetada que sea representativa del Universo de *Twitter*.

Etapa 1 de evaluación de desempeño

Para evaluar la precisión de la base actual de usuarios chilenos es necesario etiquetar manualmente si efectivamente la persona clasificada es chilena o no. Como la cantidad de personas de esta base es muy alta, se construye una muestra representativa que permita obtener la precisión de toda la base con un cierto nivel de confianza. Considerando el número de población de 1.638.317, un nivel de confianza de 99 % y un error de 3 %, se requiere que se evalúe una muestra de al menos 1.842 personas para que el resultado de la precisión sea representativa con respecto al total.

Dado lo anterior, se extraen aleatoriamente 2.000 usuarios de la base de 1.638.317 usuarios chilenos recopilados por la heurística. Se etiqueta manualmente cada registro para identificar si el usuario vive en Chile o en otro país. La etiqueta de país se realiza bajo el supuesto de que todos son chilenos a menos que la información de su perfil en su conjunto (campo *location*, descripción, zona horaria) o los *tweets* demuestren lo contrario.

Según lo anterior, se identificó que en la muestra de 2.000 registros, 15 usuarios no vivían en Chile actualmente. Por esto, se determina que la precisión de la heurística es de 99,25 %. Extrapolando los resultados a la base total, esto indica que aproximadamente 12.300 usuarios

de dicha base están clasificados como chilenos pero no lo son.

Etapa 2

En base al mismo set de validación que se utiliza para el modelo país, se analiza el desempeño de la heurística asumiendo que dicha base de usuarios sería el Universo de *Twitter*. Para esto, mediante un código de Python se identifica si en el campo *location* de los usuarios de la base aparecen ciertas palabras que contiene la heurística (Apéndice C). En caso de que aparezcan, entonces el usuario es clasificado como chileno y si no, queda clasificado como extranjero.

En la Tabla 5.3 se observan los resultados que se obtienen tras analizar la clasificación de usuarios chilenos mediante la misma metodología de la heurística en la base de validación del modelo país.

Clases	Heurística			
	Prec. (%)	Rec. (%)	F1-Score (%)	Support
Chile	74,13	33,72	46,35	510
Otro	95,71	99,21	97,43	7616
Accuracy	95,10 %			8126

Tabla 5.3: Desempeño de la heurística en la base de validación modelo país

Como se aprecia en los resultados expuestos y comparando con lo que se obtiene con el modelo país (Tabla 4.9), dicho modelo tiene mejor desempeño que la heurística en la clasificación de chilenos en la base de validación. Como las clases están desbalanceadas no se compara *Accuracy*, pero al observar el *F1-Score* de 'Chile', que es la clase de interés, el modelo país supera los resultados de la heurística en un 42,72 %, el cual asciende bastante debido a que el modelo país es capaz de recopilar más usuarios chilenos que la heurística, lo que puede apreciarse en la métrica *Recall*.

Lo anterior, permite validar que el modelo país diseñado en la presente investigación supera los resultados obtenidos por la heurística tanto a nivel de *precision* como a nivel de *recall*, siendo capaz de clasificar mejor y a una mayor cantidad de usuarios chilenos gracias a la incorporación del aprendizaje automático basado en las publicaciones de los usuarios, el lenguaje que utilizan en la interfaz de *Twitter* y lo que ingresan en su campo *location*.

5.3. Evaluación de desempeño de modelo región

5.3.1. Análisis MNB

De la Tabla 4.14 se puede extraer que en términos de *F1-Score* Macro y Accuracy la Variación 1, es decir con las publicaciones preprocesadas, clasifica mejor que la Variación 2.

Si se compara clase a clase, 13 regiones de la Variación 1 tienen mejor F1-Score individual que en la Variación 2. En cuanto al *F1-Score* Macro mejora en un 0,4 % mientras que en Accuracy mejora 0,33 %.

Dado esto, para el algoritmo Multinomial Naïve Bayes el modelo que mejor resultados entrega es aquella en que se preprocesan los *tweets* y *retweets* que publican los usuarios (Variación 1).

5.3.2. Análisis SGD

De la Tabla 4.16 se observa que, al igual que con Multinomial Naïve Bayes, la Variación 1 tiene mejor *F1-Score Macro* y *Accuracy* que la Variación 2. Esta mejora se traduce en un 0,19 % y 0,52 % respectivamente.

Por lo tanto, en el caso de Stochastic Gradient Descent, la Variación que arroja mejores resultados es aquella en que las publicaciones son preprocesadas. De todas formas, la diferencia entre ambas variaciones no es tan amplia pero se puede concluir que los *tweets* y *retweets* preprocesados ayudan a a realizar mejores clasificaciones por región.

5.3.3. Análisis RF

De los valores indicados en la Tabla 4.18 se desprende que al igual que en los algoritmos anteriores, aquella Variación que arroja mejores resultados es la Variación 1. Por ende, en los tres algoritmos utilizados para el modelo región se valida cuantitativamente el aporte del preprocesamiento en la tarea de clasificación, aunque se insiste en que la mejora no sobrepasa el 1 % por lo que no es tan notoria.

En el caso de Random Forest, sólo la 7ma región tiene una *F1-Score* más bajo en la Variación 1 que en la Variación 2, pero todas las demás regiones arrojan una métrica *F1-Score* similar o mejor que la Variación 2.

5.3.4. Análisis General

Tras comparar los modelos entrenados mediante tres algoritmos diferentes y con dos variaciones (Tablas 4.14, 4.16 y 4.18), se obtiene finalmente que el modelo que mejor realiza la tarea de clasificación por región a los usuarios de *Twitter* que viven en Chile es el algoritmo Stochastic Gradient Descent utilizando como *features* la matriz Tf-idf de las palabras extraídas de los *tweets* y *retweets* publicados por cada usuario en el periodo de un año luego de ser preprocesados mediante diferentes técnicas.

Para dicho modelo, si se realiza un análisis por región se observa que las precisiones más altas se obtienen en las regiones extremas, es decir, las regiones de más al norte o más al sur. Como lo son la 15, 1, 2, 10, 11 y 12. Esto se puede atribuir a que es más fácil identificar

habitantes de dichas regiones dado que tienen lenguajes más particulares que en la zona más central de país. De todas formas, la región 7 también arroja buenos resultados de precisión, aunque no es extrema.

A continuación se muestran resultados tras utilizar el modelo entrenado obtenido mediante validación cruzada en el Capítulo anterior en datos desconocidos para este. La base de validación para el modelo región cuenta con 5.503 usuarios chilenos, tal como se muestra en la Tabla 5.4.

Clases	Modelo			
	Prec. (%)	Rec. (%)	F1-Score (%)	Support
1	55,04	55,49	55,27	364
2	50,46	44,32	47,19	370
3	57,69	38,46	46,15	312
4	56,22	33,42	41,92	365
5	49,36	23,66	31,99	486
6	43,72	33,44	37,89	323
7	51,59	38,58	44,14	337
8	16,21	68,55	26,22	372
9	44,79	50,00	47,25	344
10	44,50	47,56	45,98	349
11	75,45	51,55	61,25	322
12	55,44	47,02	50,89	336
13	26,81	17,02	20,82	523
14	52,67	36,26	42,95	353
15	52,46	36,89	43,32	347
Prec. Macro	47,87			5503
Rec. Macro	40,40			
F1 Macro	41,78			
Accuracy	40,40			

Tabla 5.4: Validación de resultados modelo región SGD

De los resultados expuestos anteriormente se observa un *F1-Score* de 41,78 %, el cual es un valor bastante cercano al obtenido mediante validación cruzada en el capítulo anterior al momento de entrenar el modelo y se encuentra dentro del intervalo de confianza que se obtuvo en la Tabla 4.16 ya que solo es 0,76 % menor que el valor promedio. En cuanto a *Accuracy*, este también se encuentra dentro del intervalo al 95 % de confianza y es sólo 0,33 % más bajo con respecto al valor promedio. Esto indica que el modelo entrenado se comporta de acuerdo a lo esperado sobre datos desconocidos, por lo que las métricas obtenidas no presentan un sesgo considerable por sobreajuste o subajuste. Entonces, se valida el desempeño del mejor modelo encontrado para la clasificación por región de usuarios chilenos.

Analizando la matriz de confusión obtenida en la validación del modelo región, registrada en la Tabla 5.5, se observa que la principal equivocación del modelo se concentra en la clasificación errónea a la clase 8, es decir, la VIII Región del Bío-Bío. Todas las regiones

tienen más de 40 clasificaciones equivocadas asociadas a la 8va región. Curiosamente, en la matriz de confusión que resulta de las predicciones tras la validación cruzada (Anexo A), la principal equivocación era en la clase 13, es decir, Región Metropolitana, no la 8. Esto podría deberse a que aproximadamente el 50 % de la población en Chile se concentra en ambas regiones, según el Censo 2017, por lo que para el modelo entrenado es más fácil confundir entre dichas regiones con una que tiene menos habitantes. Esto ya que posiblemente el usuario no haya utilizado palabras que induzcan al modelo a clasificarlo como a la región real, y lo clasifica a una región más 'genérica' como el Bío-Bío o la Metropolitana.

Por otro lado, la clase 11, es decir, la Región de Aysén, es la que menos equivocaciones tiene y ninguna región tiene 10 o más clasificaciones erróneas asociadas a la clase 11. Esto se puede deber a que es la región que menos habitantes tiene según el Censo 2017, por lo que aquellos usuarios que fueron considerados en la base de entrenamiento como habitantes de la Región de Aysén representan bastante bien a su región en cuanto a contenido de publicaciones según palabras que utilizan.

		Predicha														
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Real	**1**	202	12	11	8	5	7	8	60	7	10	4	9	5	9	7
	2	18	164	5	6	8	8	9	93	13	7	3	8	15	4	9
	3	10	15	120	13	8	12	8	45	6	18	3	10	20	8	16
	4	9	18	10	122	7	8	8	98	21	14	4	12	12	6	16
	5	13	19	5	8	115	17	14	185	28	19	5	13	30	8	7
	6	13	10	10	5	8	108	19	62	19	15	7	8	28	6	5
	7	10	5	3	3	9	8	130	87	25	9	3	4	25	9	7
	8	14	10	5	6	10	7	6	255	9	11	1	11	16	5	6
	9	7	10	1	3	8	6	7	79	172	13	4	3	18	11	2
	10	10	6	6	8	6	10	1	66	11	166	5	11	15	14	14
	11	12	3	7	10	11	7	6	46	17	3	166	7	13	9	5
	12	9	12	6	7	8	10	1	63	12	15	2	158	17	8	8
	13	11	14	9	6	13	12	16	284	18	17	2	16	89	10	6
	14	12	10	1	4	10	17	10	79	21	27	7	4	15	128	8
	15	17	17	9	8	7	10	9	71	5	29	4	11	14	8	128

Tabla 5.5: Matriz de confusión validación modelo región

En cuanto a los resultados obtenidos por el modelo región entrenado (Tabla 4.16), se observa que mediante este sólo se logra clasificar aproximadamente el 40,73 % de los registros, lo que es menos de la mitad del total. Este valor si bien es bajo para poder determinar la región de gran parte de la población de Chile que utiliza *Twitter*, si se considerara una clasificación aleatoria, cada usuario tendría una probabilidad de 0,067 de ser clasificado correctamente a la región que pertenece debido a que son quince clases. Esto se traduce en que dicho 'modelo' aleatorio tendría un *Accuracy* de 6,67 %. Por lo tanto, el hecho que el modelo región logre un *Accuracy* de 40,73 % +/- 3,36 % mejora bastante dicho escenario.

Además, cabe destacar que actualmente el Web Intelligence Centre no cuenta con ningún método que pueda identificar la región de los usuarios chilenos de *Twitter*, por lo que aunque el modelo clasifique correctamente a menos de la mitad de los prospectos, es de utilidad como

punto de partida para el análisis tanto a nivel del proyecto SONAMA como de OpinionZoom.

Capítulo 6

Conclusiones

En el presente capítulo se exponen las conclusiones de la investigación realizada en los capítulos anteriores. Posteriormente se indican las limitaciones de este trabajo de título junto con propuestas de trabajo futuro para profundizar lo desarrollado.

6.1. Conclusiones

El trabajo que se realiza en el presente informe se hace bajo el contexto de dos proyectos del Web Intelligence Center del Departamento de Ingeniería Industrial de la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile. Estos son el proyecto SONAMA y el proyecto OpinionZoom, los cuales buscan caracterizar a los usuarios de *Twitter* a partir de la geolocalización de manera de poder determinar dónde viven a nivel regional si es que son de Chile.

Para lograr lo anterior se realizan dos modelos de clasificación con técnicas de Machine Learning para, primero, identificar a los usuarios chilenos de todo el Universo de *Twitter*, y segundo, identificar en qué región de Chile viven dichos usuarios. Para diseñar y entrenar los modelos requeridos, lo primero que se realiza es la construcción de bases de datos de entrenamiento para cada modelo en las que se combinan 3 formas de identificación del país en que vive un usuario, por lo que es capaz de representar de mejor forma el comportamiento de los usuarios de *Twitter*.

Los objetivos planteados inicialmente se cumplieron al llevar a cabo en primer lugar una revisión bibliográfica sobre lo que hasta el momento se ha realizado con respecto a la geolocalización en diferentes granularidades para así después generar una metodología tanto para la construcción de bases de entrenamiento como para los modelos a diseñar y los algoritmos que se prueban. De esto resulta el modelo país y el modelo región que permiten llevar a cabo la tarea de clasificación a nivel de Chile-No Chile y luego a nivel de regiones de Chile.

En el ámbito de la clasificación de usuarios chilenos en el Universo de *Twitter*, al utilizar el conjunto de *tweets* y *retweets* preprocesados de los usuarios junto con lo que ingresan en el

campo *location* y el lenguaje de su interfaz de *Twitter*, se logran bastantes buenos resultados para identificar a los usuarios chilenos tanto a nivel de la precisión de los clasificados como de la cantidad recopilada en función del total existente. El buen desempeño del modelo país encontrado se atribuye en gran parte al aporte de las variables de contexto utilizadas en el entrenamiento, ya que tal como se analizó en el Capítulo anterior, el no uso del conjunto de publicaciones del usuario disminuía levemente los resultado de desempeño. Esto quiere decir que tanto el campo *location* como el lenguaje son atributos en que los usuarios chilenos logran diferenciarse con usuarios del resto del mundo y mediante aprendizaje automático se logra capturar dicha diferencia. Esto podría indicar que existen ciertas *features* extraídas de dichas variables, ya sean palabras en particular en el campo 'location' o el lenguaje utilizado, que juntas dan buenos indicios sobre la clasificación del usuario. Esto se valida al probar el modelo entrenado en una base construída sólo por usuarios geolocalizados. Que si bien los resultados obtenidos en cuanto al *F1-Score* de 'Chile' decaen en un 8 % aproximadamente, la clasificación para usuarios chilenos continúa siendo alta. De hecho, que dicha métrica haya caído más que lo que se esperaba con el intervalo de confianza, podría deberse a que justamente con los usuarios etiquetados por geolocalización sea más difícil asociarlos a Chile mediante sus publicaciones y el lenguaje.

Aún así, el modelo país entrenado logra superar los resultados que se obtienen con la heurística que actualmente utiliza el Web Intelligence Center para obtener una base de usuarios chilenos. Al analizar el desempeño de esta, se obtuvo que dentro de los usuarios que la heurística recopila como chilenos, la precisión es de un 99,25 %, lo cual es bastante alta. Pero, al comparar el desemepeño del modelo país y de la heurística en la misma base, el modelo país supera los resultados tanto a nivel de precisión como de recall, por lo que se concluye que el modelo país planteado en el presente trabajo de título es de utilidad para generar una base de usuarios chilenos más amplia y similar en precisión. Sin embargo, esta base sólo estaría compuesta por usuarios que hayan publicado algún *tweet* o hayan *retweeteado* la publicacion de otro usuario.

En cuanto al modelo región, no se obtienen resultados que permitan clasificar correctamente a la mayoría, es decir sobre el 50 %, pero considerando un *F1-Macro* de 42,54 % y *Accuracy* de 40,73 % tras validación cruzada de 5-iteraciones se puede decir que el modelo es mejor que la situación actual, por lo que es útil de utiliza como *baseline* y un punto de partida en la inferencia de un lugar más granular que el país para los usuarios de *Twitter* que viven en Chile. Alguna de las hipótesis de los resultados son las siguientes:

1. No existe una diferencia significativa entre la forma de expresarse por *Twitter* entre los usuarios de las diferentes regiones

2. Debido a que *Twitter* es una plataforma de opinión y como tiene la posibilidad de *retweetear*, el contenido viene principalmente de ciertas regiones y se difunde a lo largo de todo el país, por lo que existe una tendencia a homogeneizar la forma de expresarse por *Twitter* y no hay espacio para diferenciar en ese sentido

3. En la misma línea del punto anterior, para entrenar el modelo se utiliza el conjunto de *tweets* y *retweets*. Como los *retweets* son contenido de otro usuario de *Twitter*, entonces incorporar dichas publicaciones ensucia los datos que se quieren identificar para poder

clasificar por región en cuanto a las publciaciones. Es por esto que, para validar la hipótesis de que efectivamente no existe alguna correlación entre las palabras que se utilizan en los *tweets* y la región en la que la persona vive, se debe entrenar un modelo que no incorpore *retweets*.

Por lo tanto, para el modelo región es posible validar la hipótesis de investigación ya que se logra identificar que el conjunto de publicaciones es un aporte para la tarea de clasificación, aunque no suficiente. Posiblemente, si se incorporaran variables de contexto, como campo location, los resultados serían mejores, pero en ese caso la construcción de la base de entrenamiento no podría basarse en el campo location como se realizó en esta investigación.

La presente investigación en cuanto a la predicción de usuarios chilenos dentro del universo de *Twitter* y de la región de Chile en que vive puede ser un gran aporte para nuestro país si es que se implementa. En particular, para el caso del proyecto SONAMA, poder comenzar a trabajar con estos modelos en una plataforma que pueda ir monitoreando cada cierto periodo de tiempo (semanal, mensual) el consumo de marihuana y alcohol por región y a nivel país, según lo que se *tweetea*, podría ser una herramienta que permita ir evaluando en el corto plazo las políticas públicas que se han implementado, los lugares públicos en que el Gobierno debe estar alerta, poder identificar los focos geográficos en que hay mayor consumo, entre otros.

La Encuesta Nacional de Drogas y Alcohol actualmente arroja resultados sobre el ascenso/descenso mensual del consumo, por lo que utilizando las publicaciones de los usuarios de *Twitter* diferenciándolas para usuarios chilenos y por región, en base a lo investigado en el presente informe, dicha métricas podría irse evaluando in situ y no esperar 2 años para verificar si de un mes a otro el consumo disminuyó. Por esto, se sabría con tiempo si es que necesario realizar ciertas modificaciones a nuevas normas y/o políticas públicas implementadas.

Además, si se implementa y logra buenos resultados en conjunto con el SENDA, sería un paso más para el impacto en el mundo que el Web Intelligence Centre desea lograr, siendo actores relevantes en el área de Data Science y sus aplicaciones en la salud.

Finalmente, se concluye que para este tipo de investigaciones relacionadas al aprendizaje de máquinas, además de considerar las variables más relevantes para ser utilizadas en el modelo, es clave dedicar una gran parte del tiempo a la construcción de la base mediante la cual se entrenará el modelo y con la que se validará. Las consideraciones que se tienen al momento de construir estas bases son fundamentales para evitar los sesgos y posible sobreajuste a los datos de los modelos entrenados, por lo que la rigurosidad y consciencia de los posibles sesgos que éstas puedan tener es clave.

6.2. Trabajo futuro

A partir de los resultados expuestos y las conclusiones obtenidas, se propone a continuación posibles líneas de trabajo para continuar profundizando en cuanto a la geolocalización para caracterizar de una manera más completa a los usuarios de las redes sociales. Para esto, se

propone:

- **Modelos sólo con variables de contexto que incluyan tanto usuarios activos como no activos:**
  Con respecto al modelo país, se recomienda analizar si sólo mediante variables de contexto se puede clasificar a usuarios chilenos considerando tanto a usuarios activos como no activos de *Twitter*. Esto ya que al exponer los resultados de las diferentes variantes de los tres algoritmos que se probaron para el modelo país, se rescata que para los tres algoritmos la Variación que sólo consideraba variables de contexto obtenía métricas levemente menores que las demás variaciones por lo que sería interesante estudiar si dicho comportamiento es extrapolable a los usuarios no activos de manera que se pueda obtener una base más completa de usuarios chilenos que no solo considere a los activos.

- **Enfoque en la red de contactos de *Twitter* para el modelo región:**
  En cuanto a los modelos a realizar, la presente investigación se enmarcó en modelos que consideran información del usuario a nivel de contenido y de contexto, pero tal como se comenta en el Capítulo 2, existe una tercera rama de la información que se puede extraer de la plataforma de *Twitter* que es la red de contactos.

  Dado esto, para el caso del modelo región podría ser una buena opción para experimentar la clasificación por región en base a la red seguidores y seguidos que se forma en *Twitter* y que permitiría realizar análisis de redes sociales basadas en grafos. Si bien estas redes igual estarían afectadas por el enfoque de opinión y noticiero que tiene *Twitter*, podría ser que dicha característica sea menos relevante que al estudiar el contenido que publican los usuarios directamente.

- **Análisis de palabras características de cada región:**
  Si bien al diseñar el modelo región se consideran las palabras como tokens, estos sólo eran un atributo más dentro del modelo que aportaban el valor indicado por la matriz Tf-idf construída. El hecho de incluir la matriz Tf-idf, aquellas palabras que tienen una baja frecuencia se eliminan de la matriz según el threshold que se defina. Por esto, aquellos tokens que puedan ser muy característicos de alguna región quedan fuera del modelo. Por esto, se propone realiza un análisis para encontrar un conjunto de tokens que sea característico para cada región de manera que el modelo identifique apariciones de estos tokens en particular y mediante éstos se realiza la tarea de clasificación. Esta técnica es llamada 'palabras indicativas', que son aquellas que caracterizan a un lugar en particular.

- **¿Geolocalización en Twitter?:**
  Los datos personales y los términos y condiciones de las aplicaciones están siendo cada vez más exigentes, y *Twitter* no se queda atrás. Si bien esta plataforma comenzó siendo como una de las más propensas para ser estudiadas por el nivel de datos públicos que contenía y ser muy rica en datos, cada vez se está restringiendo más y prueba de esto es la eliminación de la zona horaria y del UTC-Offset de la API REST.

  Como OpinionZoom está muy enfocado en opinión en tiempo real y sobre todo en reclamos, *Twitter* sigue siendo la red social que contiene más información con respecto a esto. Es por esto que si bien es conveniente comenzar con el estudio de otras redes,

poder canalizar los esfuerzos del estudio de *Twitter* sólo en base a las publicaciones y/o redes de seguidores y seguidos sería de gran importancia para el estudio de la geolocalización para no depender de variables consideradas más personales, como el campo *location* o la misma zona horaria. En el caso de SONAMA, por el foco de este proyecto posiblemente la información de *Twitter* vaya siendo cada vez menos útil para la tarea de geolocalización dada las nuevas restricciones que esta red está teniendo y el foco en la opinión.

Bibliografía

[1] "Estudio del comportamiento de los Chilenos en Twitter | / AnaliTIC." [Online]. Available: https://www.analitic.cl/blog/estudio-chilenos-en-twitter

[2] "History of the Web." [Online]. Available: https://webfoundation.org/about/vision/history-of-the-web/

[3] "Stochastic Gradient Descent — scikit-learn 0.20.0 documentation." [Online]. Available: http://scikit-learn.org/stable/modules/sgd.html

[4] "¿Cuáles son las cuentas de Twitter más influyentes de Chile?" [Online]. Available: http://noticias.universia.cl/cultura/noticia/2016/07/22/1142043/cuales-cuentas-twitter-influyentes-chile.html

[5] O. Ajao, J. Hong, and W. Liu, "A Survey of Location Inference Techniques on Twitter," *J. Inf. Sci.*, vol. 41, no. 6, pp. 855–864, Dec. 2015. [Online]. Available: http://dx.doi.org/10.1177/0165551515602847

[6] O. Ajao, D. P, and J. Hong, "Location Inference from Tweets using Grid-based Classification," Jan. 2017. [Online]. Available: https://arxiv.org/abs/1701.03855

[7] M. P. Andrioletti, "Detección y monitorización del consumo y consumo de riesgo de alcohol en usuarios chilenos Twitter," *Repositorio Académico - Universidad de Chile*, 2017. [Online]. Available: http://repositorio.uchile.cl/handle/2250/146741

[8] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[9] Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: a content-based approach to geo-locating twitter users," in *Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM '10*. Toronto, ON, Canada: ACM Press, 2010, p. 759. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1871437.1871535

[10] L. Chi, K. H. Lim, N. Alam, and C. J. Butler, "Geolocation Prediction in Twitter Using Location Indicative Words and Textual Features," p. 8.

[11] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep. 1995. [Online]. Available: http://link.springer.com/10.1007/BF00994018

[12] V. Cortés, "Diseño e implementación de un sistema para monitorear el consumo y opinión sobre la marihuana en Twitter," *Repositorio Académico - Universidad de Chile*, 2016. [Online]. Available: http://repositorio.uchile.cl/handle/2250/141030

[13] M. Davara Rodríguez, *Anuario de Derecho de las Tecnologías de la Información y las Comunicaciones (TIC) 2004*. Fundación VODAFONE, Madrid, 2004.

[14] N. Donges, "The Random Forest Algorithm," Feb. 2018. [Online]. Available: https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd

[15] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," *AI Magazine*, vol. 17, no. 3, p. 18, 1996.

[16] Y. G. Garc, "Algoritmos SVM para problemas sobre big data," p. 68.

[17] Hau-wen Chang, Dongwon Lee, M. Eltaher, and Jeongkyu Lee, "@Phillies Tweeting from Philly? Predicting Twitter User Locations with Spatial Word Usage." IEEE, Aug. 2012, pp. 111–118. [Online]. Available: http://ieeexplore.ieee.org/document/6425775/

[18] B. Hecht, L. Hong, B. Suh, and E. H. Chi, "Tweets from Justin Bieber's Heart: The Dynamics of the Location Field in User Profiles," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '11. New York, NY, USA: ACM, 2011, pp. 237–246. [Online]. Available: http://doi.acm.org/10.1145/1978942.1978976

[19] W. Huang, I. Weber, and S. Vieweg, "Inferring nationalities of Twitter users and studying inter-national linking." ACM Press, 2014, pp. 237–242. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2631775.2631825

[20] S.-B. Kim, K.-S. Han, H.-C. Rim, and S. H. Myaeng, "Some Effective Techniques for Naive Bayes Text Classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 11, pp. 1457–1466, Nov. 2006. [Online]. Available: http://ieeexplore.ieee.org/document/1704799/

[21] Y. Kim, J. Huang, and S. Emery, "Garbage in, Garbage Out: Data Collection, Quality Assessment and Reporting Standards for Social Media Data Use in Health Research, Infodemiology and Digital Disease Detection," *Journal of Medical Internet Research*, vol. 18, no. 2, p. e41, Feb. 2016.

[22] F. Laylavi, A. Rajabifard, and M. Kalantari, "A Multi-Element Approach to Location Inference of Twitter: A Case for Emergency Response," *ISPRS International Journal of Geo-Information*, vol. 5, no. 5, p. 56, Apr. 2016. [Online]. Available: http://www.mdpi.com/2220-9964/5/5/56

[23] Y. Lechevallier and G. Saporta, Eds., *Proceedings of COMPSTAT'2010*. Heidelberg: Physica-Verlag HD, 2010. [Online]. Available: http://link.springer.com/10.1007/978-3-7908-2604-3

[24] M. Litvak, N. Vanetik, E. Levi, and M. Roistacher, "What's up on Twitter?

Catch up with TWIST!" in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*. Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 213–217. [Online]. Available: http://aclweb.org/anthology/C16-2045

[25] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. Rosenquist, "Understanding the Demographics of Twitter Users," in *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, vol. 11, Jan. 2011.

[26] Mozilla, "World Wide Web," Dec. 2016. [Online]. Available: https://developer.mozilla.org/es/docs/Glossary/World_Wide_Web

[27] A. Noulas, S. Scellato, N. Lathia, and C. Mascolo, "Mining User Mobility Features for Next Place Prediction," in *in Location-based Services," in Data Mining (ICDM), 2012 IEEE 12th International Conference on. IEEE, 2012*, pp. 1038–1043.

[28] T. E. Reiten, "Classification with Multiple Classes using Naïve Bayes and Text Generation with a Small Data Set using a Recurrent Neural Network," Ph.D. dissertation, University of Agder, 2017.

[29] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors," p. 10, 2010.

[30] Sang-Bum Kim, Kyoung-Soo Han, Hae-Chang Rim, and Sung Hyon Myaeng, "Some Effective Techniques for Naive Bayes Text Classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 11, pp. 1457–1466, Nov. 2006. [Online]. Available: http://ieeexplore.ieee.org/document/1704799/

[31] A. Schulz, "A Multi-Indicator Approach for Geolocalization of Tweets," p. 10.

[32] Q. Yuan, G. Cong, Z. Ma, A. Sun, and N. M. Thalmann, "Who, where, when and what: discover spatio-temporal topics for twitter users," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '13*. Chicago, Illinois, USA: ACM Press, 2013, p. 605. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2487575.2487576

[33] X. Zheng, J. Han, and A. Sun, "A Survey of Location Prediction on Twitter," *arXiv:1705.03172 [cs]*, May 2017. [Online]. Available: http://arxiv.org/abs/1705.03172

Apéndice A

		Predicha														
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Real	**1**	658	46	29	31	76	39	32	42	70	29	21	27	158	48	35
	2	50	578	66	43	104	42	34	47	98	37	16	38	204	35	41
	3	26	46	531	74	107	55	29	38	68	43	28	61	183	33	67
	4	31	39	57	549	112	74	24	42	89	46	24	56	208	34	51
	5	25	35	47	49	566	50	48	73	176	45	34	59	413	44	51
	6	34	36	60	47	120	500	49	59	63	50	41	46	169	50	57
	7	42	28	42	39	90	55	558	93	86	30	21	34	171	45	27
	8	24	40	30	36	160	51	60	553	159	41	36	33	334	38	30
	9	28	30	32	35	99	37	32	39	666	37	24	32	188	50	28
	10	25	33	67	28	91	52	32	39	80	537	38	57	174	40	66
	11	29	21	24	29	77	45	26	44	67	48	738	43	113	48	28
	12	22	43	57	50	86	51	15	32	58	45	33	631	158	26	55
	13	40	48	43	50	316	86	65	104	272	57	39	58	884	73	38
	14	35	25	44	31	125	46	16	51	103	59	34	42	173	519	45
	15	49	28	65	55	104	55	24	40	74	56	27	54	186	38	496

Tabla A.1: Matriz de confusión validación cruzada modelo región SGD

Apéndice B

Campo location	cantidad
Santiago de queretaro - Mexico	1
méxico área metropolitana	1
méxico y san antonio, tx	1
san miguel de tucuman	15
grand casablanca, marruecos	15
coronel de marina leonardo ros, argentina	15
concepcion del uruguay, argentina	16
san miguel de tucuman argentina	280
san carlos	275
san miguel de tucuman	218
melbourne, victoria	212
santiago, dominican republic	206
casablanca,usa	182
san antonio, tx	169
san carlos de bariloche, argen	114
san carlos de bariloche	105
santiago de compostela	96
santiago de cali bogota	94
victoria, tamaulipas mexico	93
santiago del estero	79
santiago, nuevo león mexico	76
san fernando del valle de cata argentina	72
san fernando, argentina	67
santo tomé, argentina	67
san miguel, peru	65
la victoria, peru	59
san antonio, texas	58
san martín de los andes	56
san miguel de tucuman arg	54
san antonio de los altos	52

Tabla B.1: Ejemplos de campos location que fueron clasificados como de Chile por la heurística

Apéndice C

‘chile’ ‘arica’ ‘parinacota’ ‘tarapaca’ ‘antofagasta’ ‘atacama’ ‘coquimbo’ ‘valparaiso’
‘metropolitana’ ‘higgins’ ‘maule’ ‘bio’ ‘araucania’ ‘los lagos’ ‘los rios’ ‘aysen’
‘magallanes’ ‘maipu’ ‘puente alto’ ‘la florida’ ‘vina del mar’ ‘las condes’ ‘san bernardo’
‘temuco’ ‘rancagua’ ‘penalolen’ ‘puerto montt’ ‘santiago’ ‘concepcion’ ‘pudahuel’
‘talca’ ‘la serena’ ‘quilicura’ ‘coquimbo’ ‘la pintana’ ‘nunoa’ ‘iquique’ ‘curico’
‘chillan’ ‘talcahuano’ ‘valdivia’ ‘el bosque’ ‘quilpue’ ‘copiapo’ ‘osorno’ ‘recoleta’
‘calama’ ‘renca’ ‘villa alemana’ ‘cerro navia’ ‘providencia’ ‘punta arenas’
‘la granja’ ‘conchali’ ‘estacion central’ ‘chiguayante’ ‘colina’ ‘ovalle’ ‘macul’
‘los andes’ ‘melipilla’ ‘coronel’ ‘linares’ ‘pedro aguirre cerda’ ‘quinta normal’
‘lo espejo’ ‘lo barnechea’ ‘san pedro de la paz’ ‘lo prado’ ‘alto hospicio’
‘san joaquin’ ‘la reina’ ‘san miguel’ ‘quillota’ ‘penaflor’ ‘san antonio’ ‘huechuraba’
‘hualpen’ ‘san ramon’ ‘vitacura’ ‘la cisterna’ ‘lampa’ ‘cerrillos’ ‘talagante’
‘san fernando’ ‘buin’ ‘independencia’ ‘padre las casas’ ‘san felipe’ ‘paine’
‘concon’ ‘coyhaique’ ‘rengo’ ‘villarica’ ‘tome’ ‘calera’ ‘penco’ ‘vallenar’
‘san carlos’ ‘angol’ ‘castro’ ‘padre hurtado’ ‘lota’ ‘san vicente”limache’
‘machali’ ‘ancud’ ‘constitucion’ ‘arauco’ ‘san javier’ ‘molina’ ‘cauquenes’
‘san clemente’ ‘puerto varas’ ‘parral’ ‘la union’ ‘la ligua’ ‘chimbarongo’
‘lautaro’ ‘panguipulli’ ‘calbuco’ ‘curanilahue’ ‘maipo’ ‘canete’ ‘victoria’ ‘el monte’
‘nueva imperial’ ‘rio bueno’ ‘quellon’ ‘illapel’ ‘graneros’ ‘monte patria’ ‘shile’
‘chilito’ ‘santiasko’ ‘santiasco’ ‘scl’ ‘sch’ ‘stgo’ ‘pta arenas’ ‘casa blanca”casablanca’

Tabla C.1: Palabras que identifica la heurística para clasificar a usuarios chilenos a partir del campo location