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

DETECCIÓN Y MONITORIZACIÓN DEL CONSUMO Y CONSUMO DE RIESGO DE ALCOHOL EN USUARIOS CHILENOS DE *TWITTER*

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERA CIVIL INDUSTRIAL

MARÍA PÍA ANDRIOLETTI MÉNDEZ

PROFESOR GUÍA:
JUAN DOMINGO VELÁSQUEZ SILVA

MIEMBROS DE LA COMISIÓN:
FELIPE ESTEBAN VILDOSO CASTILLO
CARLOS FRANCISCO IBÁÑEZ PIÑA

SANTIAGO DE CHILE
2017

RESUMEN DE LA MEMORIA PARA OPTAR AL
TITULO DE: Ingeniera Civil Industrial
POR: María Pía Andrioletti Méndez
FECHA: 25/05/2017
PROFESOR GUIA: Juan Domingo Velásquez Silva

DETECCIÓN Y MONITORIZACIÓN DEL CONSUMO Y CONSUMO DE RIESGO DE ALCOHOL EN USUARIOS CHILENOS DE *TWITTER*

El consumo de alcohol es algo totalmente normalizado en nuestra sociedad. Es famosa la frase del escritor Charles Bukowski "Si ocurre algo malo, bebes para olvidar, si ocurre algo bueno, bebes para celebrarlo; y si no pasa nada, bebes para que pase algo". El alcohol está presente en cada una de las celebración y forma parte importante de nuestras tradiciones, sin embargo, es el primer factor de riesgo que causa más muerte y discapacidad en Chile, tiene una alta prevalencia en los accidentes de tránsito y se asocia a la violencia y los delitos, esto se traduce en millonarios costos tanto a nivel monetario como sociales, en el año 2006 se estimó el costo que genera el consumo en un 1,14 % del PIB de Chile. Estas consecuencias negativas resultan de gran preocupación para las instituciones de salud que buscan hacer frente a este problema de salud pública.

La explosión de las redes sociales se presenta como una alternativa factible para poder obtener información relevante a un bajo costo, comparado con otros métodos para monitorizar el comportamiento de la población. Realizar una encuesta implica costos y tiempo: por eso, este trabajo surge de la necesidad de la Unidad de Adicciones del Departamento de Psiquiatría y Salud Mental del Hospital Clínico de la Universidad de Chile de realizar una monitorización en tiempo real del consumo de alcohol.

La hipótesis de investigación de este trabajo plantea que es posible identificar el consumo de alcohol en la población y establecer la existencia de consumo de riesgo de alcohol a partir de la información disponible en *Twitter* y el contenido generado en esta red social.

El objetivo principal de esta memoria es diseñar una aplicación utilizando herramientas de *Text Mining*, *Data Mining*, *Social Network Analysis* y *Sentiment Analysis*, que permitan identificar y cuantificar la prevalencia del consumo de alcohol en la población chilena, así como también la existencia de consumo de riesgo de esta droga, utilizando la información generada en *Twitter* por usuarios chilenos, y verificar que esta información extraída refleja el comportamiento de la población general en materia de alcohol.

Para modelar el consumo de alcohol se diseñaron cuatro modelos. La *precision* para el caso de la clase de interés para cada uno de los modelos fue: 0,842 para el modelo de consumo de alcohol en *tweets*, 0,977 para el modelo de mención de políticas en *tweets*, 0,860 para el de consumo de alcohol en usuarios y 0,600 para el de consumo de riesgo.

La principal conclusión de este trabajo es que se comprueba la hipótesis de investigación. Los resultados obtenidos reflejan el comportamiento de la población general en materia de consumo y consumo de riesgo de alcohol y son comparables con la Encuesta Nacional de Drogas. Los modelos mostrados son capaces de modelar y replicar la información obtenida de esta encuesta.

“I choose to have faith, because without that, I have nothing... It’s the only thing that’s keeping me going”.
— Michael Scofield

Agradecimientos

A mi dulce y amado Agustín, muchas gracias por llenar mi vida de tu infinita alegría y tu amor inocente, por tus palabras sinceras en los momentos precisos, por hacerme reir, por los hermosos dibujos que me regalas, por tus cálidos abrazos, tus besos tiernos, por ser mi héroe, mi razón de vivir, mi estrella más luminosa en medio de la oscuridad y mi único amor verdadero e incondicional, sin ti nada de esto tendría sentido. Gracias por elegirme como tu mamá.

A Eugenio, por el apoyo incondicional, su preocupación, su cariño, los libros, los sushis y chocolates, por explicarme los misterios de la ciencia con un par de lápices, tratar de entender conmigo materias que nunca vió. Por nunca abandonarme en los momentos más difíciles. Por creer siempre en mis capacidades. Por revisar esta memoria. Gracias por motivarme a mejorar y por todos tus consejos.

A mis padres por todo el apoyo, por cuidar y entregar tanto cariño a Agustín.

A Copito, Bronco y Cokita a quienes siempre llevaré en mi corazón y sé que me cuidan y protegen.

A Pangui, siempre estaré muy agradecida por todo lo que me enseñó desinteresadamente, por explicarme con tanta paciencia cuando nada me funcionaba. A las personas del WIC que me ayudaron en este trabajo: Gaspar, Joaquín, Sebastián, Yerko, Romi, Pipe, Nicole, Felipe V. A al profesor Juan Velásquez por permitir que realizara mi memoria en el WIC y al psquiatra Carlos Ibañez por su asesoría en el desarrollo de esta memoria. A Lalo por los memes, las imágenes y videos de animales, por el apoyo emocional y la confianza, por recomendarme nuevas canciones para hacer más agradable la realización de este trabajo.

A Wentworth Miller, por ser mi fondo de pantalla, por la nueva temporada de Prison Break y por mostrarme que los sueños se hacen realidad, solo necesitamos tener un poco de fe.

A las personas que hacen posibles páginas como Google, Starkover y todas los sitios que me ayudaron a solucionar muchos errores del código. Infinitas gracias a estos héroes sin capa que comparten desinteresadamente sus conocimientos y que fueron tremendamente útiles durante mi vida universitaria, una parte importante de lo que he aprendido ha sido gracias a internet.

Tabla de Contenido

Índice de tablas

Índice de figuras

Capítulo 1

Introducción

El presente capítulo entrega al lector una presentación al trabajo de investigación. Primero, se muestran los antecedentes generales para contextualizar la problemática abordada, para luego describir el proyecto y su justificación. Luego, se plasman el objetivo general y los objetivos específicos. Después, se da a conocer la hipótesis de investigación que sustenta el proyecto, la metodología a utilizar, junto a los resultados esperados y los alcances de este trabajo. Finalmente, se detalla la estructura de este informe.

1.1. Contexto de Trabajo

El presente trabajo se desarrolla dentro del marco del proyecto CORFO código 13IDL2-23170 titulado "Opinion Zoom: Plataforma de análisis de sentimientos e ironía a partir de la información textual en redes sociales para la caracterización de la demanda de productos y servicios". Este proyecto involucra la exploración de grandes bases de datos disponibles gratuitamente en la red con el fin de recopilar, organizar y extraer nuevo conocimiento, el cual debe ser traducido a información valiosa, capaz de ayudar a la toma de decisiones por parte de los organismos interesados.

Este trabajo nace a partir del interés del psiquiatra Carlos Ibáñez, de la Unidad de Adicciones del Departamento de Psiquiatría y Salud Mental del Hospital Clínico de la Universidad de Chile, en encontrar herramientas complementarias a las actualmente existentes en el área de la salud mental, con el fin de disminuir las tasas de consumo de drogas tanto lícitas como ilícitas, y en este caso particular, de alcohol. A partir de esta necesidad, se deriva el proyecto SONAMA, que consiste en una plataforma que permite realizar seguimiento o monitorización en línea y en tiempo real de la opinión y consumo de marihuana y alcohol, y dependencia de este último, en los usuarios chilenos de Twitter.

1.2. Antecedentes Generales

1.2.1. Chile y el consumo de alcohol

A nivel Latinoamericano, Chile es el país que lidera el ranking de consumo per cápita de alcohol, según el estudio titulado "Global Status Report on Alcohol and Health 2014"[52] de la Organización Mundial de la Salud (OMS), publicado en el año 2014. En él se señala que el promedio de consumo de alcohol en el país, considerando la población mayor a quince años de edad, es de 9,6 litros de alcohol puro per cápita. Esto se contrasta con el promedio de la región de América que es 8,4. Si se considera solamente a la población que consume alcohol, en el caso de los hombres el consumo per cápita es de 19,2 litros de alcohol puro, y en el caso de las mujeres es de 9,3 litros de alcohol puro.

Otro de los indicadores que usa la OMS para monitorear los efectos del consumo de alcohol en el mundo es la *Escala de años de vida perdidos*, en inglés *Years of Life Lost (YLL) Score*. Esta es una escala con valores de 1 a 5 la cual es calculada en base al porcentaje de años de vida perdidos que pueden ser atribuibles al alcohol, donde 1 es el porcentaje menor y 5 es el porcentaje más alto [53] [55]. En el caso de Chile este índice es 5. La prevalencia de episodios de consumo excesivo de alcohol en el caso de los consumidores masculinos es 13,5 % y en el caso de los consumidores femeninos es 0,1 %, lo que indica que el consumo problemático de alcohol se da mayormente en la población masculina. En el año 2010, el porcentaje promedio de trastornos por uso de alcohol[1] considerando ambos sexo era de 5 %, Y en el caso de la población masculina y femenina, esta métrica es de 8,5 % y 1,5 %, respectivamente.

Por otra parte, de acuerdo al estudio nacional de Carga de Enfermedad y Carga Atribuible a Factores de Riesgo en Chile, el consumo de alcohol es la primera causa de Años de Vida Saludables Perdidos, llamados AVISA (Figura 1.1). En Chile, el consumo de alcohol como factor de riesgo, se relaciona con el 12,4 % de lo años de vida saludables (AVISA) perdidos por muerte o discapacidad. Esto corresponde al doble de los AVISA producido por obesidad (6.3 %) o por presión arterial (5.6 %).

El uso de alcohol es un tema de especial atención para la unidad de adicciones dado que está asociado a innumerables externalidades negativas para la sociedad, además de graves enfermedades tanto físicas como mentales para quien lo consume. El uso y abuso de alcohol en Chile es un tema relevante ya que se estima que alrededor del 10 % de las muertes a nivel nacional se pueden atribuir al consumo de alcohol. Esta cifra implicaría unas 9.500 muertes anuales, es decir, cada día mueren aproximadamente 26 chilenos en quienes su causa de muerte estuvo relacionada el consumo de alcohol [46].

Las consecuencias económicas y sociales derivadas del consumo excesivo y dependencia al alcohol son frecuentemente muy difíciles de determinar ya que es causa de una gran variedad de enfermedades o agravamiento de las mismas. Por otra parte, es también motivo de accidentes automovilísticos, de violencia, de rupturas de parejas, de muchos casos de violencia intrafamiliar, e incluso puede provocar la muerte. Además, el tratamiento de los problemas asociados al alcohol deben ser abordados desde diferentes perspectivas, mediante un equipo multidisciplinario de espe-

[1] Incluyendo la dependencia del alcohol y el uso nocivo del alcohol.

cialistas y acompañándose de tratamiento farmacológico que generalmente es costoso. Gran parte de los costos atribuibles al alcoholismo son de carácter indirecto y con un impacto muy elevado sobre la sociedad. En numerosos casos de homicidios y suicidios el alcohol ha estado presente como protagonista.

Según [47], el consumo de alcohol contribuye de manera importante a las tres principales causas de muerte (lesiones no intencionales que también reciben el nombre de accidentes, homicidio y suicidio) entre las personas de 12 a 20 años en Estados Unidos.

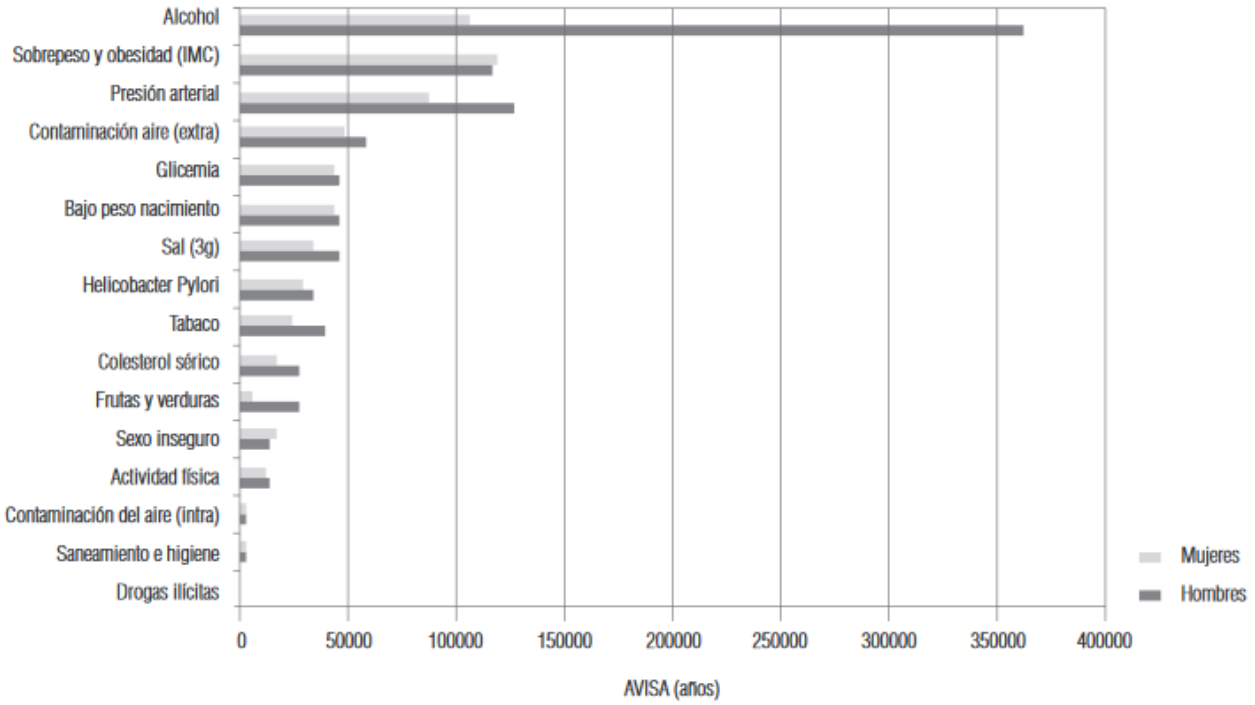

Figura 1.1: Carga de AVISA (años) atribuible a Factores de Riesgo según género, Chile 2007.

Fuente: Estudio de carga de enfermedad y carga atribuible a factore de riesgo en Chile, 2007. Imagen extraída de [46].

1.2.2. Relación entre alcohol y suicidio

Existe un amplio espectro de factores de riesgo del suicido, donde uno de ellos es el uso nocivo de alcohol. Está documentado que el consumo y el abuso de alcohol esta altamente ligado a al comportamiento suicida [40]. Las personas que se han suicidado tienen niveles positivos de concentración de alcohol en la sangre [67].

Chile muestra un preocupante deterioro de la salud mental, lo que se ve evidenciado en que Chile se ubica en el cuarto lugar dentro de los países con mayor tasa de suicidio en la región de América: en el periodo comprendido entre los años 2005 a 2009, la tasa era de 11,68 personas por cada 100 mil habitantes. En Chile, en 1990, esta tasa era de 5,6 por cada 100 mil habitantes, lo que indica que la tasa ha ido en aumento en las últimas dos décadas [20]. En la región de América Latina y el Caribe, se observaron incrementos en las tasas de suicidio totales y para cada sexo

durante el periodo de 20 años entre 1990 y 2009. Es decir, Chile es parte de esta tendencia al alza del número de suicidios en la región [20].

En Chile, el cambio en la tasa de suicidio fue de +54,9 % entre los años 1995 y 2009, mientras que en los países de la OECD, el cambio en la tasa de suicidio fue de -10,3 % en el mismo periodo. Es decir, mientras los países de la OECD, en promedio, disminuyeron su tasa de suicidio en 15 años, en Chile esta tasa aumentó. Entre los años 2000 y 2011 la tasa de cambio en los países de la OCDE fue de -7,0, mientras que en Chile fue de +19,8, lo que lo ubica en el tercer puesto de los países con porcentajes de cambio en la tasa de suicidio más alta.

La Organización Mundial de la Salud tiene un plan para la prevención del suicidio cuyo objetivo es reducir la tasa de suicidio en un 10 % en los países para el 2020 [54]. Por ello, la OMS plantea que una de las intervenciones relevantes que se deben tomar, es la creación de políticas para reducir el uso nocivo de alcohol. Por lo tanto es importante buscar alternativas que permitan monitorear el consumo de alcohol en la población.

1.2.3. Importancia de las redes sociales e Internet

A continuación se presentan algunas estadísticas de Twitter para mostrar la importancia que tiene este medio en nuestra sociedad actualmente.

- 500 millones de personas visitan Twitter cada mes sin iniciar sesión
- Existe un total de 1,3 billones de cuentas, de las cuales 320 millones son usuarios activos
- Cada día 500 millones de Tweets son enviados, es decir cada segundo son enviados 6000 Tweets.[2]
- En Chile su presencia alcanza 1700000 usuarios diarios (Talento Virtual, 2015)

1.2.4. Web Intelligence Centre

El Web Intelligence Centre (WIC) es un centro de investigación perteneciente a la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile. Su principal enfoque es la inteligencia web y la utilización de herramientas de *Data Science* para resolver problemas. Una de las tareas que se realizan es tratar de aplicar los conocimientos de *Data Science* a otras áreas, como por ejemplo medicina o los negocios.

En el WIC se realiza investigación aplicada en el área de Web Intelligence, la cual se encuentra respaldada por numerosas publicaciones en revistas científicas internacionales. La docencia es otro foco del centro, donde se dictan cursos de orientación práctica acerca de las Tecnologías de Información y su aplicación en los negocios.

El WIC es parte del Web Intelligence Consortium[3], el cual agrupa distintos centros de investigación y desarrollo en Inteligencia Web alrededor de todo el mundo.

[2] https://www.brandwatch.com/2016/05/44-twitter-stats-2016/ visitada 15 de septiembre, 2016

[3] `http://wi-consortium.org/`, visitado el 27 de Diciembre del 2016

En el sitio web del WIC[4] se puede encuentra plasmada su misión, su visión y los objetivos, y son los que se muestran a continuación:

Misión

Desarrollar investigación de frontera en el campo de Tecnologías de Información creando nuevas soluciones para abordar problemas complejos de ingeniería utilizando herramientas basadas en la Web de las Cosas.

Visión

Ser un líder a nivel internacional en la investigación de tecnologías de información y comunicaciones aplicadas a la resolución de problemas del mundo real.

Objetivos

- Publicar en las principales revistas, conferencias y editoriales relacionadas con Web Intelligence.
- Proveer un 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.

1.3. Descripción del Proyecto y Justificación

Las empresas gastan enormes cantidades de recursos para conocer en profundidad a sus clientes. Generalmente, se crean segmentos que permiten caracterizar cada grupo de clientes para así poder entregar los productos o servicios ideales para ellos. Además, deben enfocar sus esfuerzos en los clientes que tienen una mayor disposición a adquirir sus productos o servicios, que es lo que lleva a esas personas a preferir cierta marca por sobre otra y como evoluciona en el tiempo. La empresa tiene incentivos a querer conocer a sus compradores y potenciales clientes con el fin de maximizar su beneficio neto. Esta idea, usada y conocida en las empresas privadas, puede ampliarse y adaptarse a contextos sociales. En el caso de la Unidad de Adicciones del Departamento de Psiquiatría y Salud Mental del Hospital Clínico de la Universidad de Chile, el objetivo es poder caracterizar y conocer a los consumidores de alcohol para poder idear políticas públicas eficaces para desincentivar el consumo de riesgo de alcohol en la población y también conocer la opinión de ellos en torno al alcohol.

Las consecuencias negativas producto del consumo de alcohol es de preocupación para Unidad de Adicciones del Departamento de Psiquiatría y Salud Mental del Hospital Clínico de la Universidad de Chile y otras instituciones de salud, debido a lo enormes costos asociados al fenómeno del

[4] `http://wic.uchile.cl/`, visitado el 27 de Diciembre del 2016

alcohol en el país. En [50] se concluyó que el costo conjunto para el país que genera el consumo de drogas ilícitas y alcohol en Chile, en el año 2006, fue de $882.512 millones de pesos (en moneda de valor de ese año). En particular, para el alcohol los costo estimado fueron alrededor de $550.000 millones, estos representan un 1,14 % del Producto Interno Bruto de Chile del 2006. Los mayores costos son representados por las pérdidas de productividad por los años de vida saludables perdidos (AVISA) y por pérdidas en accidentes automovilísticos [50].

Debido a la consecuencias resultantes de consumo nocivo de alcohol existen distintas alternativas para hacer frente a este problema. Un nuevo enfoque para la prevención del consumo de alcohol se encuentra mediante el uso de *Social Media*. En [63], se explora el uso problemático de alcohol utilizando la red social de *Facebook*.

Existe una gran cantidad de información disponible en las redes sociales, la cual es generada en tiempo real por los usuarios. En estas plataformas ellos comentan sus intereses, actividades diarias y publican información personal de alto valor, lo que las transforman en importantes fuente de información para analizar. A partir de esta oportunidad, surge la necesidad de poder extraer de forma automática la información disponible en la Web. Esto es posible a través de las técnicas de *Opinion Mining* o también conocidas bajo el nombre de Análisis de Sentimientos, términos que hacen referencia a tratamiento computacional de las opiniones, los sentimientos y la subjetividad presentes en un texto [57].

En varios estudios se han usado técnicas de minería de datos y text mining para la información disponible en las redes sociales. Por esta razón surge como una alternativa la utilización de estas técnicas en materia de drogas. Anteriormente, esto se realizó con la marihuana [18] dando buenos resultados.

1.4. Objetivos

El presente trabajo busca lograr los siguientes objetivos:

1.4.1. Objetivo general

Diseñar una aplicación, usando herramientas de *Sentiment Analysis*, que permita identificar y cuantificar la prevalencia del consumo de alcohol en la población chilena, la existencia de consumo de riesgo, así como también las percepciones que tiene la población frente a esta droga, utilizando la información generada en *Twitter* por usuarios chilenos.

1.4.2. Objetivos específicos

1. Estudiar el estado del arte actual relacionado con el análisis de consumo de alcohol dentro del contexto de *Opinion Mining*, además de los conceptos teóricos dentro del área de psiquiatría necesarios para entender el fenómeno.

2. Definir una metodología basada en *Sentiment Analysis* para analizar el contenido extraído de *Twitter* para estudiar el consumo de alcohol en la población.
3. Extraer patrones de comportamiento en los datos extraídos de *Twitter* y definir métricas para la caracterización de la prevalencia de consumo de alcohol en la población chilena y de los niveles de uso problemático y de riesgo de esta sustancia.
4. Analizar y evaluar los resultados obtenidos con el fin de generar conclusiones y probar el desempeño de los métodos utilizados.

1.5. Hipótesis de investigación

La hipótesis de investigación de este trabajo plantea:

Es posible identificar el consumo de alcohol en la población a partir del contenido generado en la red social Twitter y establecer la existencia de consumo de riesgo de alcohol.

En otras palabras, lo que se hará en este trabajo es analizar si la información que los usuarios generan en *Twitter* es suficiente para ser usada como fuente para detectar el consumo de alcohol y si existe un consumo excesivo de esta sustancia, a nivel agregado.

Para ello, se utilizarán herramientas del área de las tecnologías de la información conocida por su nombre en inglés como *Sentiment Analysis* (también conocida como *Opinion Mining*) y *Text Mining*, que se refiere al uso del procesamiento del lenguaje natural, análisis de textos y lingüística computacional para descubrir patrones comunes utilizando una gran cantidad de bases de datos.

1.6. Metodología

Para alcanzar los objetivos planteados anteriormente en este punto se propone la siguiente metodología.

El primer paso será realizar una documentación con respecto a las dos áreas con las que está relacionado este trabajo: por un lado, lo que respecta al área de psiquiatría vinculada al uso de sustancias psicoactivas, específicamente al consumo de alcohol y los factores de riesgo que explican su consumo y abuso y por otro lado, se estudiará el estado del arte en materia de técnicas de *Text Mining*, *Opinion Mining* y *Data Mining* para evaluar su aplicación en el presente proyecto. En este punto también se investigará la aplicación de éstas técnicas en las redes sociales en materia de adicciones.

El segundo paso será la identificación y definición del problema que se planea resolver para el cliente.

En tercer lugar se recolectarán datos desde *Twitter*. Para esto, se utilizarán las bases de datos existentes en el centro WIC que han sido construidas a lo largo de la existencia de este centro; también se utilizarán las API de *Twitter* para extraer información actualizada y reciente y también

información histórica.

Posteriormente, con los datos recolectados se definirán las palabras claves que servirán para encontrar la información necesaria en *Twitter*.

Se definirán las *features* o variables que serán utilizadas para construir los modelos de clasificación, primero para categorizar los *tweets* y a continuación las variables necesarias para modelar el consumo de alcohol y el consumo de riesgo y dependencia en la población chilena.

Se construirán los algoritmos de clasificación y regresión, con los cuales se podrá obtener los patrones existentes entre los consumidores y consumidores de riesgo de alcohol.

Se construirán métricas de rendimiento para medir la precisión de los algoritmos creados en el punto anterior.

Los algoritmos serán validados con el psiquiatra para asegurar que cumplen con los requerimientos del cliente y son una herramienta útil para él.

1.7. Resultados esperados

El presente proyecto pretende alcanzar los siguientes resultados:

- La construcción de un corpus que abarque los criterios de clasificación determinados del estudio y que permitan caracterizar el consumo del alcohol.
- Una base de datos que almacene información relacionada con el consumo y abuso de alcohol.
- Modelos predictores que entreguen información sobre el uso y consumo de riesgo de alcohol en *Twitter*.
- Una aplicación que permita procesar el contenido obtenido desde *Twitter* para adquirir información orientada al interés del cliente.

1.8. Estructura del informe

A continuación, se detallará un plan de trabajo a partir de la metodología descrita anteriormente, el que permitirá llevar un orden del trabajo realizado.

1. El primer capítulo trata sobre la introducción a la memoria, tomando en cuenta el contexto, los objetivos de la memoria y la metodología seguida.
2. El segundo capítulo describe el marco teórico y conceptual usado en el presente trabajo, se trata las dos partes que tiene esta memoria, tanto la parte de psiquiatría ligadas al consumo de sustancias y las herramientas necesarias para construir el modelo.
3. El tercer capítulo trata sobre el estudio de predictores que ayuden a la detección de consumo de alcohol en las redes sociales, específicamente *Twitter*.

4. Diseño: trata sobre los pasos a seguir para poder recolectar la información necesaria en *Twitter* y los modelos a utilizar.
5. Implementación: este capítulo trata de como fue llevada a cabo el diseño detallado en el capítulo anterior.
6. Resultados: en este capítulo se analizan los resultados obtenidos y si se lograron los objetivos propuestos.
7. Conclusiones y trabajo a futuro.

Capítulo 2

Marco Teórico

En esta sección se acerca al lector a los conceptos fundamentales y necesarios para lograr un buen entendimiento del desarrollo de este proyecto.

2.1. World Wide Web

La World Wide Web (WWW) fue propuesta por Tim Berners-Lee en 1990. Al principio, la web 1.0 se utilizaba para mostrar información de sólo lectura mediante páginas estáticas, con actualizaciones periódicas a cargo de un Webmaster.

En el año 2004 fue definido el concepto de Web 2.0, por Dale Dougherty. Esta "nueva versión" sugiere un enfoque distinto al anterior, ya que el usuario pasa a ser el centro, creando una web dinámica en donde los usuarios pasan a ser los principales creadores de contenido.

2.1.1. Web 2.0

En [16] se detallan las principales categorías de la Web 2.0 presentadas a continuación

Blog: Este es un sitio web en donde las personas expresan sus ideas, pensamientos y comentarios. Las entradas de los blogs, también llamadas post, se basan en contenido propio expuesto en forma de bitácora, mostrados, generalmente, en orden cronológico, en primer lugar los más recientes.

Really Simple Syndication (RSS): este es un formato XML utilizado para resumir la información y links de recursos para compartir el contenido de blogd o páginas web. El objetivo es informar a los seguidores sobre actualizaciones de blogs o sitios web en los que el usuario está interesado.

Wikis: estas son páginas webs que son editables fácilmente para cualquier persona que acceda

a ellas. Se caracterizan por tener un lenguaje y estructura simples y a veces pueden estar centrada en algún tema particular.

Comunidades: en estos sitios web se organizan en torno a compartir información acerca de contenido en particular. En estas comunidades se comparten videos, fotos, enciclopedias públicas o social bookmarking.

Foros: son sitios para intercambiar ideas e información sobre intereses específicos.

Redes Sociales: Estos son un grupo online de aplicaciones que conectan a personas según sus intereses e información que comparten. En ellas se crean relaciones entre los usuarios, que pueden ser mutuas o no. Actualmente están disponibles las aplicaciones para estar conectado a través de los celulares.

2.2. Proceso Knowledge Discovery in Databases (KDD)

Knowledge Discovery in Databases (KDD), conocido en español como Descubrimiento de Conocimiento a partir de Bases de Datos, es el proceso organizado de identificación de patrones válidos, nuevos, útiles y comprensibles a partir de grandes y complejos conjuntos de datos [45]. Fayyad et al [25] lo define como el "proceso no trivial de identificación de patrones válidos, originales, útiles y entendibles acerca de la data". Data se define como un conjunto de hechos, y patrón es una expresión en algún lenguaje que describe un subconjunto de la data o un modelo aplicable al subconjunto. El proceso KDD es un método interactivo e iterativo que consiste de seis pasos, definidos como:

Comprender el dominio de la aplicación: Primero que todo, es fundamental analizar y comprender aquello que se hará con el proceso, tomando como centro los objetivos del usuario final y el entorno en el que se llevará a cabo el descubrimiento del conocimiento. A medida que el proceso avance, este análisis podría redefinirse.

Seleccionar y crear el set de datos para el proceso: Consta de la extracción y selección de datos a utilizar desde una base de datos, la cual dependerá del tipo de problema a resolver.

Pre-procesamiento de datos: Consiste en la limpieza de los datos, detectando los valores fuera de rango, faltantes o nulos.

Transformación de datos: Se trata de la modificación de los datos que se ingresarán al modelo.

Data Mining: Consiste en la selección apropiada de la tarea y de los algoritmos de minería de datos y su implementación para obtener los patrones subyacentes en la data que está siendo analizada.

Evaluación e interpretación: Es aquí donde se obtiene el conocimiento a partir de los datos luego de pasar por las etapas anteriores.

En la figura 2.1 se explica gráficamente el proceso.

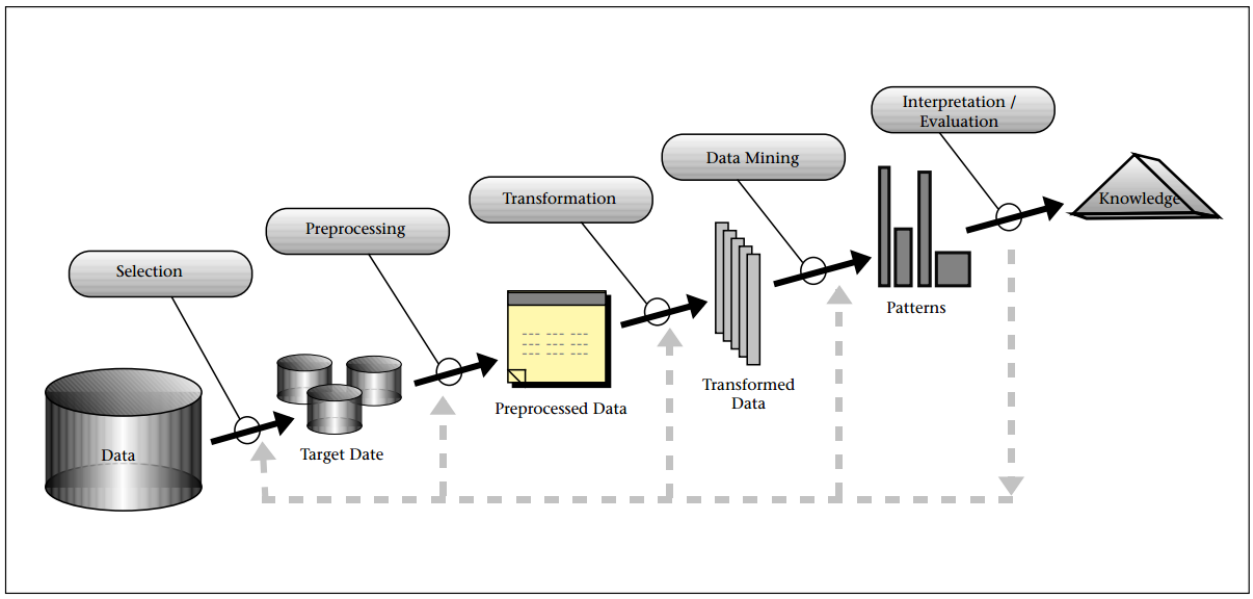

Figura 2.1: Diagrama del proceso KDD.

Fuente: Imagen extraída de [25]

2.2.1. Data Mining

Data mining se define como el proceso de descubrir nuevas correlaciones, patrones y/o tendencias a partir del análisis de grandes cantidades de datos, cuyas fuentes pueden incluir bases de datos, *Data Warehouses*, repositorios de información, datos obtenidos desde la Web o que han sido generados de forma dinámica. Es importante destacar que a pesar de que las técnicas de Data Mining permiten descubrir nuevos patrones, la utilidad que estos patrones representen serán determinados a través del proceso KDD.

2.3. Web Mining

Web Mining se define como la aplicación de las técnicas de Data Mining utilizadas para el descubrimiento y análisis de información desde la Wold Wide Web (W3). Debido al explosivo crecimiento de la Web y sus fuentes de datos, este campo ha sido objeto de especial atención debido a todas las posibilidades de descubrimiento de conocimiento que significa y los grandes desafíos futuros que eso conlleva.

2.3.1. Definición de Web Intelligence

Web Intelligence se define como el uso de técnicas avanzadas de Inteligencia Artificial y Tecnologías de Información para el propósito de explorar, analizar y extraer conocimiento del contenido web [74] . El objetivo es crear una nueva generación de productos y servicios basados en el internet.

Formalmente *Web Mining* es definido por Cooley et al en [17] como el descubrimiento y análisis de información útil desde la World Wide Web. Es posible encontrar diversos enfoques relacionados con el estudio de los diversos tipo de datos relacionados a la Web y al uso por parte de los usuarios. Se definen tres categorías que se describen a continuación detallando sus principales características.

2.3.2. Categorías

Web Usage Mining

Web Usage Mining analiza aquello que los usuarios buscan ver en internet o cualquier otro registro de su actividad, estudiando el comportamiento de éstos durante el uso de los sitios web con el objetivo de generar perfiles de usuarios. Los datos pueden ser obtenidos a través de los servidores web, proxies, aplicaciones del cliente (como navegadores web y sus historiales o cookies) y con ellos se puede obtener conocimiento del uso que los usuarios hacen de ciertos sistemas o de sus intereses. Este conocimiento es útil para ser aplicado en personalización de sistemas web, marketing, diseño, evaluación de sitios web y como un apoyo a la toma de decisiones.

Web Structure Mining

Web Structure Mining es el proceso de utilizar la teoría de gráficos para analizar los nodos (que serían las páginas) y la estructura de conexión en un sitio web (los cuales serían los nodos). Dependiendo con el tipo de los datos estructurales de la web, *Web Structure Mining* puede ser dividido en dos:

- La extracción de patrones de los hipervínculos[1] en la web. Tomando al hipervículo xomo un componen estructural que conecta la página web a una ubicación diferente.
- El minado de la estructura del documento, es decir, el analisis de la estructura como si fuera un árbol para describir el uso de etiquetas HTML o XML.

Web Content Mining

Web Content Mining es el descubrimiento de información relevante a partir de los contenidos de la Web, los cuales pueden ser textos, imágenes, audios y videos. La mayor parte de los contenidos de la Web están en formato de texto y sobre estos se aplican variadas técnicas como *Text Mining*, clústering y clasificación de texto, por nombrar algunos.

[1] En inglés hyperlink

2.3.3. Web Opinion Mining

Opinion Mining o *Sentiment Analysis*, que en español se denomina Análisis de Sentimientos, corresponde al estudio computacional de opiniones, sentimientos, subjetividad, evaluaciones, actitudes, valoración, emociones, etc. de las personas hacia entidades tales como productos, servicios, organizaciones, individuos, materias, eventos, tópicos y sus atributos expresados en un texto [2]. Ambos campos tienen algunas pequeñas diferencias, pero generalmente son utilizados como sinónimos para describir el mismo campo de estudio. Morfológicamente (o semánticamente) los términos opinión y sentimiento no son idénticos, pero a pesar de esto tanto *Opinion Mining* como Análisis de Sentimientos se enfocan en el estudio de opiniones que expresan sentimientos positivos y negativos.

Las opiniones poseen una gran importancia ya que impactan en el comportamiento de la población. Las creencias y percepciones de la realidad están fuertemente condicionadas a como otros ven dicha realidad. Generalmente, cuando alguien toma una decisión pide o busca las opiniones de los demás y toma como referencia las experiencias vividas y expresadas por otras personas. Antiguamente, el círculo de cada persona estaba delimitado por las personas más cercanas al individuo, y en el caso de las empresas y organizaciones, estas debían recurrir a métodos como encuestas, focus groups o consultorías externas.

Actualmente, con el crecimiento de los medios presentes en la W3, se puede acceder a blogs, foros de opinión y redes sociales, aumentando el poder de las opiniones. Gracias a las características de la Web 2.0, los usuarios pueden compartir con otros usuarios alrededor del mundo sus opiniones y experiencias respecto a diversos temas. Existe un gran interés de los usuarios en conocer las opiniones de los demás y en la potencial influencia que ellos pueden significar en las decisiones de otros. Esto toma cada vez una mayor relevancia para empresas y organizaciones, y prueba de esto es que existe un nicho de interesado en disponer de sistemas capaces de analizar de forma automática las opiniones generadas por clientes para poder obtener conocimiento a partir de éstas.

2.4. Técnicas de Minería de Datos

Existen varios tipos de técnicas de minería de datos, que se utilizan dependiendo del propósito y los objetivos a los que se desean llegar. Se dividen en dos grandes áreas: las orientadas a descubrimiento, para encontrar patrones y reglas, y las orientadas a la verificación, que buscan validar hipótesis [45].

Los métodos orientados al descubrimiento identifican patrones de forma autónoma y se dividen en métodos descriptivos y predictivos. Los métodos descriptivos están orientados a la interpretación de la data, cuyo enfoque se centra en el entendimiento de la forma en que los datos se relacionan con sus partes. Por otro lado, los métodos predictivos intentan crear un modelo de comportamiento de forma automática, que obtenga nuevos ejemplos y sea capaz de predecir valores de una o más variables relacionadas con el ejemplo. Además, genera patrones que facilitan el entendimiento del conocimiento descubierto. A su vez, los métodos predictivos se dividen en algoritmos de clasificación y regresión. Los algoritmos de clasificación tienen como objetivo categorizar casos futuros, mientras que los de regresión, generan un pronóstico a partir de los datos.

Los métodos orientados a la verificación, tienen como misión validar hipótesis planteadas por un agente externo. Generalmente, se ocupan métodos de estadística tradicional, como bondad de ajuste, análisis de varianzas (ANOVA), y tests de hipótesis. Esta clasificación se explica gráficamente en la ilustración 2.2.

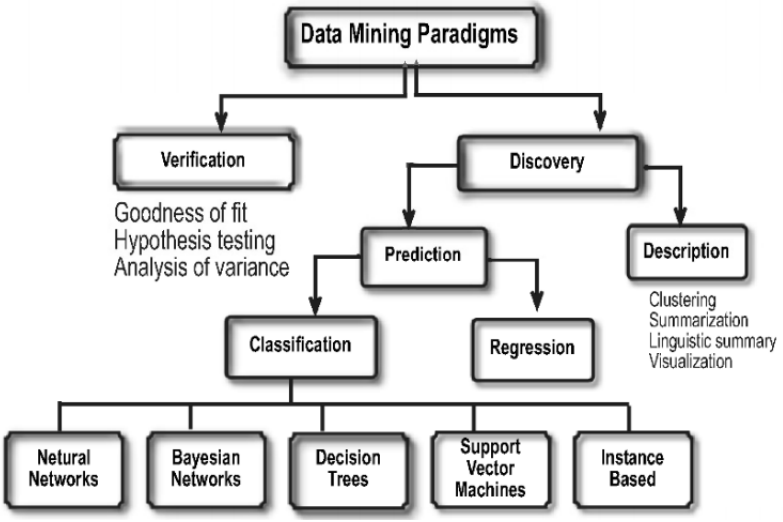

Figura 2.2: Técnicas de minería de datos.

Fuente: Imagen extraída de [45]

2.5. Text Mining

Text Mining, también conocido como *Text Data Mining* [34], es conocida como la minería de los datos de texto o el descubrimiento de bases de datos de texto. Este es un proceso en el que se extraen patrones o conocimiento útil, interesante y nuevo a partir de documentos de texto no estructurados [69]. Puede ser visto como una extensión de la minería de datos o el descubrimiento de conocimiento desde bases de datos estructuradas.

La mayor parte de la información en la web está en formato de texto. Estudios indican que el 80 % de la información en una empresa se encuentra contenida en documentos de texto [69]. Sin embargo, el *Text Mining* implica una tarea más compleja que la minería de datos, ya que los documentos de texto son inherentemente no estructurados y difusos. La minería de texto es un campo multidisciplinario, que incluye la recuperación de información, el análisis de texto, la extracción de información, *clustering*, categorización, visualización, tecnologías de bases de datos, *machine learning* y minería de datos.

2.6. Machine Learning (ML)

El concepto de *Machine Learning* se puede definir como el proceso de programar una máquina o computador para que sea capaz de entregar resultados lo suficientemente útiles en base al uso de

datos de ejemplo o experiencia pasada.

2.7. Algoritmos Supervisados y no Supervisados

Es posible distinguir los algoritmos de minería de datos entre supervisados y no supervisados. Los métodos no supervisados[2] modelan la distribución de instancias, donde estas no están previamente identificadas. Es decir, cada valor de la data no tiene asociada una etiqueta que clasifique ese valor en alguna categoría establecida. Este tipo de métodos no requiere esfuerzo de etiquetado manual, el que en algunas ocasiones podría ser costoso. En el análisis de textos los métodos más utilizados son Clustering y Topic Modeling.

Por otro lado, los algoritmos supervisados[3], conocidos también bajo el nombre de métodos de clasificación por su extendido uso para este fin, intentan descubrir las relaciones que existen entre variables o atributos independientes y una variable o atributo dependiente. Esta relación se modela como patrones de comportamiento que pueden predecir valores en base a entradas de datos [5]. En este caso entre las técnicas más popularmente utilizadas se encuentran los árboles de decisión, Suport Vector Machine, Naive Bayes Classifier, entre otras.

2.8. Toping Modeling

Probabilistic Topic Models [10] son algoritmos utilizados para descubrir los temas principales de los cuales trata una colección masiva de documentos no estructurados. Los algoritmos de Topic Modeling pueden ser adaptados a distintos tipos de datos como por ejemplo para encontar patrones en datos de genética, imágenes o en las redes sociales, de esta forma pueden tener numerosas aplicaciones. Estos algoritmos no requieren de un etiquetado previo, ya que los tópicos son derivados del análisis de los textos originales. El gran valor de Topic Modeling es que permite organizar y resumir archivos digitalizados que a gran escala sería imposible hacerlo de manera manual por un humano.

Latent Dirichlet Allocation (LDA) es el modelo más simple de topic model y es un modelo estadístico de una colección de documentos que trata de capturar la intuición que los documentos contienen múltiples tópicos

Un tópico formalmente se define como una distribución sobre un vocabulario fijo. El modelo asume que los tópicos son generados primero, antes de los documentos.

Para cada documento en una colección de documentos, las palabras son generadas en un proceso de dos etapas:

1. Elegir aleatoriamente una distribución sobre los tópicos
2. Para cada palabra en el documento:

[2]En Inglés se denomina Unsupervised Learning
[3]En inglés es llamado Supervised Learning

a. Elegir aleatoriamente un tópico de la distribución sobre los tópicos del paso 1.
b. Elegir aleatoriamente una palabra desde la correspondiente distribución sobre el vocabulario.

2.9. Redes Sociales

2.9.1. Microblogging

Microblogging es una forma relativamente nueva de comunicación, en donde los usuarios pueden compartir sus estados mediante mensajes cortos, que pueden enviar a través de sus teléfonos móviles mediante las aplicaciones para smarthphones (en español llamados teléfonos inteligentes), o a través de la Web. Estos mensajes pueden hablar sobre sus intereses, sobre su vida personal o sobre acontecimientos actuales.

El máximo exponente del *Microblogging* es *Twitter*[4] cuya popularidad ha aumentado significativamente desde que fue lanzado en Octubre de 2006 [36]. Sin embargo, han existido otras plataformas de *Microblogging*, tales como Jaiku y Pownce, pero que con el tiempo han desaparecido por diversos motivos.

2.9.2. Twitter

Orígenes

Twitter es una plataforma de *microblogging* creada en marzo del 2006 por Jack Dorsey, Evan Williams, Biz Stone y Noah Glass y lanzada en julio del mismo año. Ha tenido un crecimiento acelerado, llegando a ser el noveno sitio más visitado a nivel global [3]. Cuenta con 313 millones de usuarios activos al mes [72] y 100 millones de usuarios activos diarios [43].

Las cuentas de *Twitter* pueden representar a una persona natural, una persona famosa, una organización pública o privada, empresas o también puede ser un servicio en particular que presta una empresa. En otras palabras, existen compañías que tienen una cuenta destinada a difundir sus promociones y otra como canal de ayuda para sus clientes, etc. Un ejemplo es la empresa de Telecomunicaciones Entel, con sus cuentas @entel[5] y @entel_ayuda[6]. Por otro lado, también se puede encontrar cuentas que son una parodia de algún famoso, político o personaje de ficción.

Existe la posibilidad de verificar la cuenta, lo que permite comprobar la autenticidad de las cuentas de interés público. Esta opción existe ya que algunas veces se crean cuentas no oficiales, generalmente con usuarios del ámbito de la música, la actuación, la moda, el gobierno, la política,

[4]http://www.twitter.com
[5]https://twitter.com/entel
[6]https://twitter.com/entel_ayuda

la religión, el periodismo, los medios de comunicación, el deporte, los negocios y otras áreas de interés [73].

Elementos de *Twitter*

La red social *Twitter* cuenta con algunas características propias que se detallan a continuación.

***Tweet*:** puede ser sólo una declaración o estado escrita por un usuario, o puede ser una respuesta a otro *tweet*. Una de las restricciones es que puede contener como máximo 140 caracteres. El contenido es principalmente texto, pero también existe la opción de incluir imágenes, videos o hipervínculos.

Etiquetas o *Hashtag* (#): es una cadena de caracteres precedidos por el caracter numeral o almohadilla (#). En muchos casos, los *hashtags* pueden ser considerados como temas de actualidad, una indicación en el contexto del *tweet* o como la idea central expresada en el *tweet*. Además, los *hashtags* pueden ser adoptados por otros usuarios para contribuir con contenido similar o expresar una idea relacionada [70]. El uso masivo de un mismo *hashtag* determinará un *Trending Topic*.

Retweet (RT) : dar *retweet* es un mecanismo para citar un *tweet* de otro usuario o, en otras palabras, copiarlo en el propio perfil, señalando que el *tweet* originalmente fue publicado por otro usuario. El usuario que publicó el *tweet* original es notificado, y además el *tweet* almacena un conteo de las veces que ha sido *retweeteado* para así contabilizar su popularidad. Al *retwitearlo* existe la opción de agregar un comentario al *tweet*.

***Followers* o Seguidores:** el acto cuando un usuario se suscribe a los *tweets* de otro usuario recibe el nombre de seguir, y por lo tanto el usuario que se ha suscrito a otro es llamado seguir o *follower*. Esto significa que el *follower* recibirá en su inicio los *tweets* de los usuarios a los cuales sigue. No existe un límite de seguidores.

Following: son los usuarios que un determinado usuario sigue. No existe la necesidad que el acto de seguir o ser seguido sea recíproco como ocurre en otras redes sociales. Cada cuenta puede seguir a 5000 usuarios en total, una vez alcanzado ese límite, existe una cantidad adicional de usuarios que se puede seguir, el cual depende de la proporción entre seguidores y seguidos. Para evitar el *spam*, una cuenta de *Twitter* es técnicamente incapaz de seguir a más de 1000 usuarios al día.[7]

@ seguido de un nombre de usuario: se utiliza para mencionar a un usuario, con el fin de dirigirle directamente un *tweet* y llamar su atención.

Privacidad: los usuarios en *Twitter* pueden elegir si sus *tweets* son públicos o estén protegidos. Cuando el usuario se registra en *Twitter* sus *tweets* serán públicos de manera predeterminada. Cuando los *tweets* son públicos cualquier persona puede ver e interactuar con los *tweets*. En el caso de que los *tweets* estén protegidos, estos no aparecerán en los motores de búsqueda y solos podrán ser vistos por los seguidores que acepte el usuario, después que reciba una

[7] https://support.twitter.com/articles/73277# visitada el 2 de Octubre 2016

solicitud de un nuevo seguidor. [8]

Links: los usuarios pueden agregar a sus *tweets* enlaces (URL) a sitios externos. Estos links también pueden contener fotos. Todos los vínculos que se compartan en *Twitter* son acortados mediante el servicio http://t.co. Esto permite compartir enlaces largos manteniendo el número máximo de caracteres por *tweet*. [9] [10]

Cronología o *Timeline* (TL): se muestra los *tweets* ordenados cronológicamente de manera descendente que escriban las personas a las que sigue un usuario en particular, así como también los que ha escrito el propio usuario.

Emoticones, *emojis* y expresiones coloquiales: está permitido incorporar *emojis* a los *tweets*.

2.9.3. Interfaz de Programación de Aplicaciones (API)

La Interfaz de Programación de Aplicaciones, conocida en inglés como *Application Programming Interface* (API), es un conjunto de definiciones de subrutinas, protocolos y herramientas para la construcción de software y aplicaciones. La API es una interfaz que permite que dos programas de software puedan comunicarse entre sí.

La principal función de la API es facilitar a los desarrolladores el uso de ciertas tecnologías en la creación de nuevas aplicaciones. Las API son de gran utilidad para los desarrolladores, ya que no es necesario que el desarrollador entienda cada una de las operaciones que hace por detrás la otra aplicación para que pueda ser utilizada. Algunos de los usos de las API son la recuperación de datos o la actualización de dicha aplicación. Una de las desventajas de las API es que no están estandarizadas.

Una API puede ser dependiente o independiente del lenguaje. Una API dependiente del lenguaje puede ser utilizada sólo usando la sintaxis particular de un lenguaje de programación en el que se implementa la API. Por otro lado, una API independiente del idioma es un proveedor de servicios genéricos y puede ser llamado por cualquier idioma de programación [35].

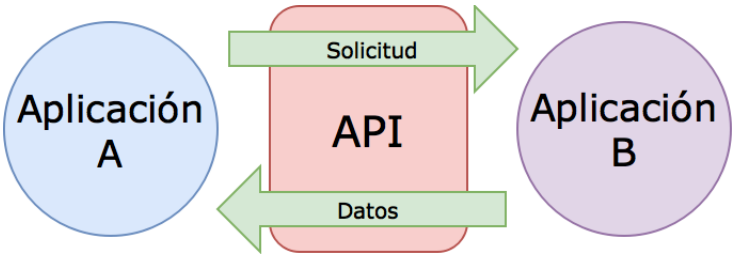

Figura 2.3: Dos aplicaciones comunicadas utilizando una API

Fuente: Imagen extraída de [12]

[8] https://support.twitter.com/articles/339960
[9] https://support.twitter.com/articles/344713#
[10] https://support.twitter.com/articles/344685#

Twitter ofrece tres API: Streaming API, Rest API y Search API, las cuales están destinadas a diferentes necesidades.

La *Streaming API* da a los desarrolladores acceso de baja latencia a un subconjunto de tweets globales que son generados en el momento en tiempo real por los usuarios de twitter. Se puede obtener una muestra aleatoria de los estados o se puede aplicar un filtro por palabras claves o por usuario. La conexión a la Stream API requiere mantener una conexión http abierta y persistente [71]. Los comandos GET, POST, and DELETE pueden ser usados para acceder a la data [9]. Se requiere de una cuenta válida de Twitter y los datos pueden ser recuperados en formato JSON.

La Rest API de Twitter proporciona acceso para leer y escribir datos en esta plataforma. Los resultados se entregan en formato JSON.

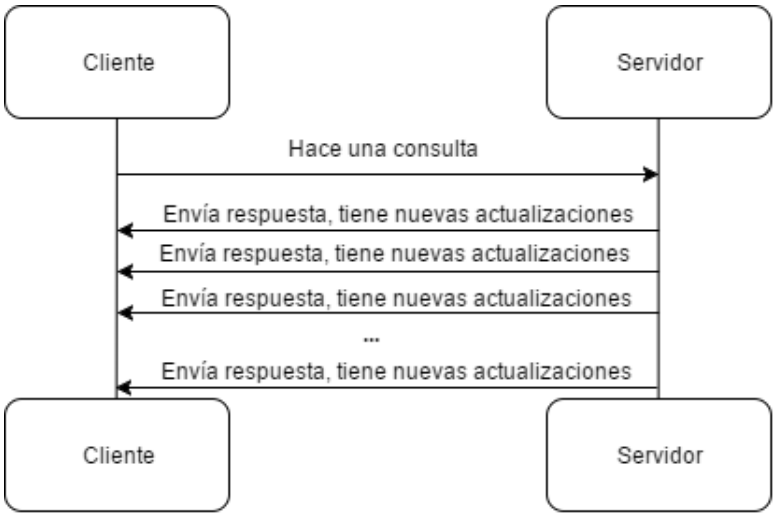

Figura 2.4: Streaming API

Fuente: Elaboración propia

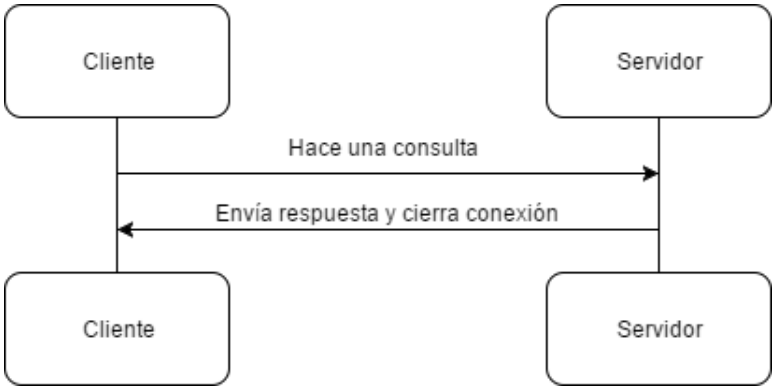

Figura 2.5: Rest API

Fuente: Elaboración propia

Con la librería Twitter4j se puede obtener la información de un usuario de Twitter que se detalla a continuación:

- La URL de la imagen de perfil
- La fecha de creación del usuario
- La descripción del usuario
- La cantidad de tweets marcados como favorito
- La cantidad de followers que tiene el perfil
- La cantidad de followings que tiene el perfil
- La ID única que twitter asigna al usuario
- El lenguaje principal que utiliza el usuario
- La cantidad de listas públicas en las que está el usuario
- La locación del usuario, en el caso que la tenga escrita
- El nombre del usuario
- El color de fondo del perfil
- La imagen de fondo en el perfil
- La URL del banner del perfil
- Los colores del perfil
- El nombre de Twitter que tiene el usuario
- El último tweet generado por el usuario
- La cantidad de tweets que ha generado el usuario
- Obtener la URL que tiene el usuario en su descripción
- Testear si el usuario ha cambiado el tema de su perfil
- Comprobar si el usuario es privado
- Comprobar si el usuario está verificado
- Testear si el usuario tiene la georeferencia de sus tweets activada
- Testear si el usuario es traductor

Por otra parte también es posible obtener la siguiente información de un tweet:

- La fecha de creación de un *tweet*
- El número de veces que el *tweet* ha sido marcado como favorito
- La geolocalización del *tweet*
- La ID del *tweet*
- El nombre e ID del usuario a quien se le hizo la respuesta si así fuera el caso
- La ID del *tweet* al que este fue respuesta
- El idioma del *tweet*
- El lugar en el caso que haya sido adjuntado
- El número de veces que el *tweet* ha sido *retweeteado*
- El texto del *tweet*
- El usuario creado por *Twitter*
- Testear si es un *retweet*
- Testear si ha ido *retweeteado*
- Comprobar si el *tweet* contiene algún contenido marcado como sensible
- Testear si el *tweet* es favorito

2.10. Extracción de información

2.10.1. Crawling

Crawling es el proceso de recopilar datos desde la Web [37]. Un *crawler*, también conocido como araña web, es un software cuyo objetivo principal es actualizar su propio contenido que almacena de ciertas web o para indexar el contenido que tiene de algunas páginas. Muchos motores de búsqueda Web y algunos otros sitios realizan este proceso para poder actualizar su propio contenido web o índices de contenido web de otros sitios web y de esta forma realizar búsquedas más rápidas y eficientes dentro de ellas.

Dado que el número de páginas en internet es extremadamente grande, incluso los rastreadores más grandes no llegan a hacer un índice completo. Por esta razón, en los años tempranos de la World Wide Web (antes del 2000), los motores de búsqueda entregaban resultados deficientes. Esto ha ido gradualmente mejorando por los motores de búsqueda modernos, ya que hoy en día entregan muy buenos resultados de forma inmediata.

En la figura 2.6 se muestra la arquitectura de un Web Crawler. Un *Crawler* además de tener una buena estrategia de rastreo debe tener una arquitectura optimizada [68]. Los *Web Crawler* son una parte central de los motores de búsqueda y los detalles sobre sus algoritmos y arquitectura se mantienen como secreto de negocio.

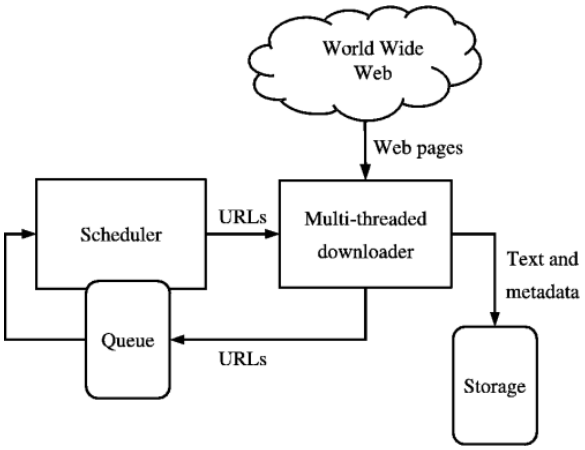

Figura 2.6: Arquitectura de un Web Crawler

Fuente: Elaboración propia

2.10.2. Preprocesamiento de datos

Tokenización

En el análisis léxico, la tokenización [29] es el proceso de romper una cadena de texto de entrada en subunidades, las cuales son llamadas *tokens* y que pueden ser palabras, frases, símbolos u otros

elementos significativos. La lista completa de tokens posteriormente se convierte en entrada para los siguientes pasos en la realización del tratamiento de Lenguaje Natural como por ejemplo el análisis sintáctico, o la para realizar *Text Mining*. La tokenización es de gran utilidad en lingüística como también en ciencias de la computación donde es parte del análisis léxico. Generalmente, los tokens están separados por caracteres de espacio en blanco, como puede ser un salto de línea, o por caracteres de puntuación. La tokenización es uno de los primeros pasos en la transformación del texto en el tratamiento de Lenguaje Natural.

Stemming

Stemming proviene del inglés *stem*. En la morfología lingüística y el campo de la búsqueda y recuperación de información (*Information retrieval*) es el proceso en el cual se lleva a las palabras a la raíz de estas. Por ejemplo, *bibliotecario* y *biblioteca*, luego de ser sometidas al proceso de *Stemming*, se transforman ambos en *bibliotec*. Esto ayuda a que posteriormente disminuya el número de palabras con las cuales trabajar para realizar la clusterización de éstas.

El algoritmo mas comúnmente utilizado para realizar el paso de *Stemming* es el algoritmo de Porter, el cual fue publicado en 1980 [59]. Lo que propone este algoritmo es eliminar sufijos de las palabras basándose en la gramática y en las reglas de reemplazo. Existen otros algoritmos de *stemming* que habían sido desarrollados antes que el algoritmo de Porter, pero éste último presenta una manera más simple y efectiva que otros métodos más complejos [59].

Porter primero definió las consonantes y las vocales, las cuales componen las palabras. Una consonante sera denotada por una (“c”), mientras que una vocal sera denotada como (“v”). una lista ccc... de longitud mayor que cero sera denotada por una (“C”) y una lista de vvv... de longitud mayor que cero será denotada por una (“V”). Cualquier palabra, o parte de una palabra, puede ser representada por alguna de estas cuatro formas:

- CVCV...C
- CVCV...V
- VCVC...C
- VCVC...V

Estos pueden ser representados en una única forma como:

$$[C]VCVC...[V]$$

En donde los paréntesis de corchete representan la presencia arbitraria de aquellas consonantes o vocales.

Usando $(VC)^m$ para denotar la presencia m veces de VC, lo anterior puede ser escrito como:

$$[C](VC)^m[V]$$

Al valor de *m* se le llama *medida*, para cada palabra o parte de palabra cuando es representada de esta manera. Cada palabra tiene su *medida*, por ejemplo, ARMADO está formado por VCVCV y posee *medida 2*, mientras que AMA está formado por VCV y su medida es 1. A continuación se

presentan otros ejemplos:

$m = 0$ $\quad QUE(CV)$, $Y(V)$, $ME(CV)$, $PROA(CV)$

$m = 1$ $\quad AMA(VCV)$, $ROMPRE(CVCV)$, $PAN(CVC)$, $NUNCA(CVCV)$

$m = 2$ $\quad ARMADO(VCVCV)$, $PRIVADO(CVCVCV)$, $PROBLEMA(CVCVCV)$

La regla que es utilizada para remover un sufijo es:

$$(\text{condición}) \; S_1 \rightarrow S_2$$

Lo anterior significa que si una palabra termina con el sufijo S_1, y la raíz antes de S_1 satisface la condición, S_1 es reemplazado por S_2. La condición está dada usualmente en términos de la medida. A continuación se muestra un ejemplo en inglés:

$$(m > 1) \; EMENT \rightarrow$$

En el ejemplo anteriormente mostrado S_1 es $EMENT$ y S_2 es nulo, por lo que si se tiene la palabra $REPLACEMENT$, está se cambiará por $REPLAC$, ya que tiene una medida de 2.

Porter propone las siguientes condiciones para la gramática inglesa:

- $*S$: La raíz termina en s
- $*v*$: La raíz contiene una vocal
- $*$d: La raíz termina en doble consonante
- $*o$: La raíz termina en *cvc*, donde la segunda c no es W, X o Y.

Porter llevó este removedor de raíces a un lenguaje llamado *Snowball*, implementándolo en diversos lenguajes como Java o C y distintos idiomas como el Español o Francés.

Lematización

Algunas veces los términos de *Stemming* y lematización son utilizados como equivalentes pero no lo son.La lematización intenta llevar a la palabra a su lema utilizando el uso de un vocabulario y del análisis morfológico de las palabras, por ello se suele juntar dos palabras que podrían ser analizadas como una sola más que obtener aquellas con la misma raíz. Por ejemplo, en inglés, *better* podría tener el mismo lema que *good*; si se hubiera realizado *stemming* estas dos palabras no estarían relacionadas.

Eliminación de Stopwords

Las *stopwords* son palabras vacías que no tienen un significado útil, o muy poco, para el análisis de documentos. Estas palabras pueden ser artículos, pronombres o preposiciones. Generalmente, estas palabras son eliminadas en la etapa de preprocesamiento de texto para así poder conservar solo los conceptos que se están analizando y entregan valor. Cabe destacar que no existe una lista

definitiva de stopwords y a veces se evita usar listas de *stop words* y no son eliminadas para realizar análisis por frases o para no tener problemas con algunas *stopwords* que tengan doble significado o presenten ambigüedad como algunos nombres. Un ejemplo de esto en español es té (sustantivo) y te (pronombre). Una alternativa es el uso de un algoritmo de stemming con el fin de reducir parte de la base lógica o dependencia de una lista de *stopwords* a filtrar.

2.11. Validación Cruzada o *Cross Validation*

La validación cruzada o *Cross Validation* es un método estadístico de evaluación y comparación de algoritmos de aprendizaje por medio de la división de datos en dos segmentos: uno es usado para aprender o entrenar un modelo y el otro es utilizado para validar el modelo [62]. Típicamente, en una validación cruzada, los set de *training* y validación son cruzados (cross-over) en rondas sucesivas de forma que cada punto en la data tiene una oportunidad de ser validado. La forma básica de validación cruzada es la *k-fold cross-validation*. Otras formas de validación cruzada son casos especiales de *k-fold cross validation* o implican repetidas rondas de *k-fold cross-validation*

Los dos objetivos en la validación cruzada son:

- Estimar el rendimiento del modelo aprendido a partir de la data disponible utilizando un algoritmo. En otras palabras, para calibrar la generalización de un algoritmo.
- Para comparar el desempeño de dos o más algoritmos diferentes y poder encontrar el mejor algoritmo para la data disponible o alternativamente para comparar el rendimiento de dos o más variantes de un modelo parametrizado.

2.11.1. K-Fold Cross-validation

En *k-fold cross-validation* la data es particionada en k segmentos de igual tamaño (o muy parecidos). Posteriormente, se realizan k iteraciones de entrenamiento y validación de manera tal que dentro de cada iteración un segmento de la data es usado para la validación, mientras que los $k-1$ segmentos restantes serán utilizados para el aprendizaje del algoritmo. La figura 2.7 muestra un ejemplo con $k = 3$. La sección oscura de la data es usada para en entrenamiento, mientras que la sección mas clara es usada para la validación. En minería de datos y *machine learning 10-fold cross-validation* ($k = 10$) es lo que comúnmente se utiliza.

Cross-validation es usado para evaluar o comparar algoritmos de aprendizaje de la siguiente forma: en cada iteración uno o más algoritmos de aprendizaje usa $k-1$ segmentos de la data para aprender uno o más modelos. Posteriormente, a los modelos se les pide que hagan predicciones sobre el segmento que fue destinado a la validación. El desempeño de cada uno de los algoritmos de aprendizaje puede ser seguido mediante el uso de determinadas métricas de desempeño, una de ellas puede ser por ejemplo, la *accuracy*. Al completarse, se tendrá para cada algoritmo k muestras de la métrica de rendimiento. Se pueden utilizar diferentes metodologías para obtener una medida agregada de estas muestras, como por ejemplo el promedio, o estas muestras pueden ser utilizadas en una prueba de hipótesis estadística para mostrar que un algoritmo es superior que otro.

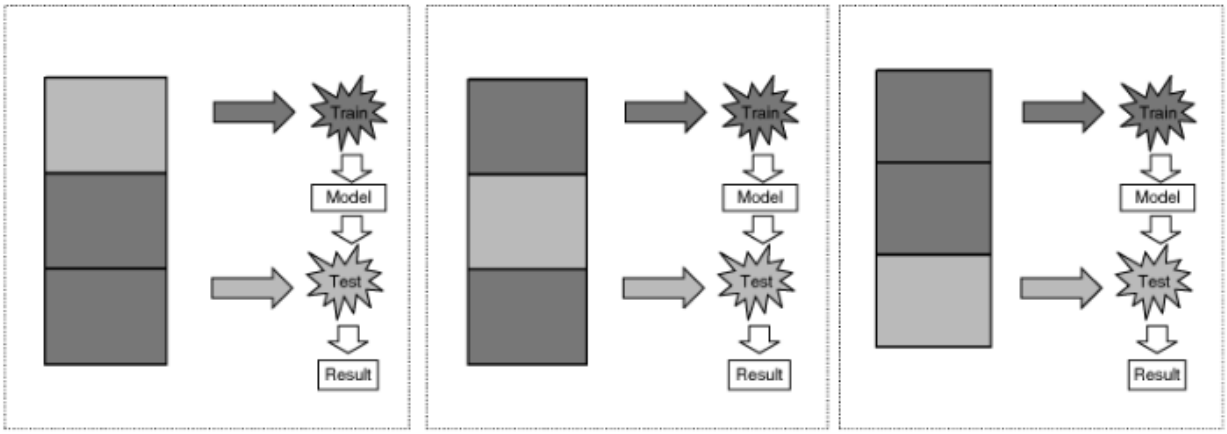

Figura 2.7: Cross-validation. Procedimiento de *three-fold cross validation*

Fuente: Imagen extraída de [62]

2.12. Kappa de Fleiss

Cohen primero introdujo el estadístico Kappa [14] y también su versión ponderada llamado *weighted kappa* [15], con el fin de medir el grado de acuerdo entre dos evaluadores que clasifican cada una de las muestras de un tema determinado en una escala nominal. El uso de *kappa* y *weighted kappa* está restringido para caso en que los evaluadores son dos y donde los mismo dos evaluadores clasifiquen cada uno de las muestras. Entonces, surge la necesidad de una generalización para los caso en que existan más de dos evaluadores y para el caso donde los evaluadores que clasifican un elemento no necesariamente son los mismos que clasifican otro.

Kappa de Fleiss es una medida estadística para evaluar la confiabilidad del acuerdo entre un número fijo de evaluadores al asignar clasificaciones categóricas a un número de artículos o elementos que son posibles de clasificar.

Kappa de Fleiss sólo se puede utilizar con valores binarios o nominales.

2.13. Evaluación de rendimiento

En esta sección se realiza la evaluación de rendimiento de los modelos de clasificación construidos. La clasificaciones hechas en este trabajo son parte de los casos típicos de clasificación por lo que han sido estudiadas en la literatura.

- ***Falso Positivo (FP)*** (o Error tipo I): corresponden a los casos predichos como positivos, pero que verdaderamente tienen un valor negativo.
- ***Falso Negativo (FN)*** (o Error tipo II): son los casos predichos como negativos, pero que en la realidad tienen un valor positivo.
- ***Verdadero Positivo (VP)*** (También conocido como éxito.): casos predichos correctamente con valor positivo.

- ***Verdadero Negativo (VN)*** (Conocido también con el nombre de rechazo correcto): casos predichos correctamente con valor negativo.

En la tabla 2.1 se muestra la matriz de confusión para el caso de clasificación binaria.

	Valor Predicho	
Valor Verdadero	Positivo	Negativo
Positivo	VP	FN
Negativo	FP	VN

Tabla 2.1: Matriz de confusión.

Fuente: Elaboración Propia

Por otro lado las métrica de evaluación de rendimiento se muestran a continuación:

1. **Precision**: esta métrica estima que la probabilidad de que una predicción del caso positivo sea correcto. Su ecuación es:

$$P = \frac{|VP|}{|VP| + |FP|} \tag{2.1}$$

2. **Recall**: es la proporción de casos que tienen un valor positivo y fueron correctamente predicho como positivos. Su ecuación se presenta a continuación:

$$R = \frac{|VP|}{|VP| + |FN|} \tag{2.2}$$

3. **F-Measure**: es una combinación entre *precision* y *recall*, donde una constante β controla el *trade-off* entre ambas métricas. La fórmula general para un número real β es:

$$F - Measure = (1 + \beta^2) \cdot \frac{P \cdot R}{(\beta^2 \cdot P) + R} \tag{2.3}$$

4. ***Area under the ROC curve (AUC)*:** La curva ROC se grafica trazando la tasa de verdaderos positivo (TPR)(también conocidos como *recall*) frente a la tasa de falsos positivo (FPR). La tasa de verdaderos-positivos también se conoce con el nombre de *recall*.

$$FPR = \frac{|FP|}{|FP| + |VN|} \tag{2.4}$$

   El indicador más utilizado en varios contextos es el área bajo la curva ROC o AUC, éste índice se puede interpretar como la probabilidad de que un clasificador ordene una instancia positiva elegida aleatoriamente más alta que una negativa elegida al azar, es decir $P(score(x^+) > score(x^-))$.

   Los modelos con altos AUCs son preferidos por sobre aquellos que presentan bajos AUCs.

Capítulo 3

Consumo de alcohol

3.1. Factores que afectan el consumo de alcohol y daños relacionados con el alcohol

A continuación se muestran una serie de factores a nivel individual y de la sociedad que afectan los patrones de consumo y que pueden incrementar el riesgo de los trastornos por consumo y otros problemas relacionados con el alcohol [6]. Existen factores ambientales, tales como el desarrollo económico, las distintas culturas, la disponibilidad de las bebidas alcohólicas, y el nivel de efectividad de las políticas relacionadas con el control del consumo y venta de alcohol, los cuales se constituyen como factores relevantes a la hora de explicar las diferencias existentes en las tendencias históricas en el consumo de alcohol y sus respectivos daños asociados a su abuso.

3.1.1. Edad

Según [32], una edad más temprana de inicio de alcohol está fuertemente relacionada con un mayor nivel de consumo abusivo a los 17-18 años de edad. Los niños, los adolescentes y las personas mayores generalmente suelen ser más vulnerable a los daños relacionados con el alcohol que otros grupos etarios, dado un determinado volumen de alcohol. Por otro lado, el inicio a temprana edad del consumo de alcohol, es decir antes de los catorce años, es un predictor del deterioro del estado de salud, ya que se asocia con un mayor riesgo de dependencia y de abuso de alcohol en edades posteriores, accidentes relacionados a vehículos motorizados, y también otros accidentes no intencionales.

Parte del mayor riesgo que se observa en jóvenes está relacionado con que, generalmente, una gran proporción del alcohol total consumido por los jóvenes se produce durante periodos de consumo intensivo. Por otro lado, los jóvenes tienen menos aversión al riesgo y pueden ser participes de conductas mientras se encuentran borrachos.

El daño relacionado con el consumo de alcohol entre las personas mayores se deben a factores un tanto diferentes de los daños relacionados con el alcohol entre los jóvenes anteriormente descritos.

El consumo de alcohol, en general, disminuye con la edad. Los adulto mayores consumen alcohol menos frecuentemente que otros grupos etarios. Por otro lado a medida que las personas envejecen, sus cuerpos suelen ser menos capaces de manejar los mismos niveles y patrones de consumo que llevaban en años atrás; esto ocasiona que las personas tengan lesiones no intencionales asociados, por ejemplo a caídas. En muchos países donde la población está envejeciendo es un tema de preocupación en materias de políticas públicas la carga de enfermedades asociadas al alcohol entre el grupo etario de mayor edad.

3.1.2. Género

El uso nocivo de alcohol es el principal factor de riesgo de muerte en varones de 15 a 59 años de edad. A pesar de esta alarmante estadística, existe evidencia que las mujeres pueden se más vulnerables a los daños asociados al alcohol para un nivel de consumo de alcohol o un determinado patrón de consumo. Esta vulnerabilidad señalada para el género femenino es un importante tema de preocupación para la salud pública, ya que se ha observado que el consumo entre las mujeres ha ido aumentando progresivamente en relación con el desarrollo económico y también asociado a los cambios de los roles de género. Un punto muy relevante es que esto puede significar gravísimas consecuencias sanitarias y sociales para los recién nacidos[1] [44] [58]

En el año 2012 el 7,6 % de todas las muertes de hombres alrededor del mundo fueron atribuibles al alcohol. En el caso de las mujeres, esta cifra fue de 4,0 %. Por otra parte, los hombres, en comparación a las mujeres, tienen una tasa mayor de carga total de enfermedades expresada en años de vida ajustados por discapacidad [1] atribuibles al alcohol; para el caso de los hombres este porcentaje es de 7,4 % comparado con el 2,3 % para las mujeres. El aumento en la carga de enfermedades entre los hombres se explica, en gran medida, porque en comparación con las mujeres, los hombres a menudo son menos abstemios, beben con mayor frecuencia y en cantidades mayores.

3.1.3. Factores de riesgo familiar

El núcleo familiar juega un rol central en el consumo temprano de alcohol y en el progreso de los problemas derivados del consumo de alcohol [48]. Una historia familiar de trastornos por consumo de alcohol se considera un factor de vulnerabilidad, tanto por razones genéticas como por razones ambientales.

Se ha observado que los trastornos asociados al uso de alcohol parental afectan negativamente la situación familiar durante la infancia.

En [66] fue comparada una muestra de 253 niños hijos de padres con problemas de alcoholismo (COAs) una muestra de 237 niños hijos de padres sin problemas de alcoholismo (non-COAs). Fueron consideradas variables tales como el uso de drogas, psicopatologías, habilidades cognitivas y personalidad. En el estudio se observó que los primeros (COAs) reportaron mayores problemas de alcohol y drogas.

[1] en inglés *disability-adjusted life years (DALYs)*

Una serie de factores de riesgo relacionados con el ámbito familiar han sido consistentemente relacionados a los problemas de alcohol en adolescentes. Algunos de estos factores de riesgo son el uso de sustancias por parte de los padres y hermanos, escasa supervisión por parte de los padres [39], conflictos familiares, ausencia de normas relativas al alcohol, y escaso apoyo y control familiar.

El consumo excesivo de alcohol en los padres no solamente predicen niveles iniciales de consumo en sus hijos, sino que además influencian el incremento de consumo a lo largo del tiempo [48].

3.1.4. Estatus socio-económico

Según [30] en el mundo desarrollado existe un mayor número de bebedores, bebedores ocasionales y bebedores con patrones de uso de bajo riesgo en los grupos socio-económicos más altos, mientras que es más común encontrar abstemios en lo grupos sociales mas pobres. Sin embargo, las personas con un nivel socio-económico bajo parecían ser más vulnerables a los problemas tangibles y las consecuencias producto el consumo de alcohol. Una de las razones que se maneja para explicar la vulnerabilidad potencialmente mayor de los grupos socio-económicos más bajos, es que estas personas son menos capaces de evitar las consecuencias negativas de su comportamiento debido a la carencia de recursos.

3.1.5. Control y regulación de alcohol

La disponibilidad de alcohol influye en el consumo de alcohol y también en su uso nocivo. En [51] se plantea la estrategia mundial recomendada por la OMS para reducir el uso nocivo de alcohol. El uso nocivo de alcohol se puede reducir si los países adoptan medidas eficaces para proteger a su población, mediante la formulación, aplicación, supervisión y evaluación de políticas públicas de reducción de uso nocivo de alcohol.

Las opciones de políticas he intervenciones se han agrupado en diez esferas de acción recomendadas, las cuales se enumeran a continuación:

1. Liderazgo, concienciación y compromiso.
2. Respuesta de los servicios de salud
3. Acción comunitaria
4. Políticas y medidas contra la conducción bajo los efectos del alcohol
5. Disponibilidad de alcohol
6. Marketing de las bebidas alcohólicas
7. Políticas de precios
8. Mitigación de las consecuencias negativas del consumo de alcohol y la intoxicación etílica
9. Reducción del impacto en la salud pública del alcohol ilícito y el alcohol de producción informal
10. Seguimiento y vigilancia.

Debido a la importancia que desempeñan las políticas en materia de control de consumo y venta de alcohol, se hace necesario incluir las percepciones de la población en esta materia.

En cuanto a las políticas referentes al alcohol presentes en Chile se encuentran las leyes que regulan el consumo y la venta o el mercado de alcohol en Chile, las cuales se detallan a continuación:

Ley sobre expendio y consumo de bebidas alcohólicas. Ley Número 19.925

En esta ley [23] se explicita la regulación que la edad mínima legal nacional para la venta de bebidas alcohólicas (cerveza, vino y destilados) es dieciocho años.

Reforma tributaria que modifica el sistema de tributación de la renta e introduce diversos ajustes en el sistema tributario. Ley Número 20.780: Impuesto a bebidas alcohólicas

En el país la venta o importación de bebidas alcohólicas paga un impuesto adicional, que se aplica sobre la misma base imponible del Impuesto a las Ventas y Servicios [19]. A continuación, se detallan las tasas de los impuestos, dependiendo de tipo de bebida alcohólica.

a. Licores, piscos, whisky, aguardientes y destilados, incluyendo los vinos licorosos o aromatizados similares al vermouth, tasa del 31,5 %.
b. Vinos destinados al consumo, comprendidos los vinos gasificados, los espumosos o champaña, los generosos o asoleados, chichas y sidras destinadas al consumo, cualquiera que sea su envase, cervezas y otras bebidas alcohólicas, cualquiera que sea su tipo, calidad o denominación, tasa del 20,5 %.

Esta norma sobre impuesto de alcoholes comenzó a regir el 1 de octubre de 2014.

Alcohol y conducción.

El Ministerio de Transporte, a través de la Comisión Nacional de Seguridad de Tránsito, CONASET, ha logrado importantes avances en materia de leyes relacionadas a alcohol y conducción.

Ley Tolerancia Cero (Ley Número 20.580): esta ley [21] entró en vigencia el 15 de marzo del año 2012 y corresponde a una modificación de la ley de tránsito número 18.290. En concreto con esta ley se disminuyó los grados de alcohol permitidos en la sangre para los conductores:

- El estado de ebriedad se estableció en 0,8 gramos de alcohol por litro de sangre. Anteriormente, era 1,0 gramos por litro de sangre.
- El estado bajo la influencia del alcohol fue fijado en 0,3 gramos de alcohol por litro de sangre. Anteriormente, este límite se encontraba en 0,5 gramos por litro de sangre.

Además fueron aumentadas las sanciones de suspensión de la licencia de conductor, dependiendo de las infracciones que éste cometa y las consecuencias de distintos niveles que pueden resultar.

Ley Emilia (Ley Número 20.770): esta ley[22] entró en vigencia en septiembre de 2014, a raíz de una petición ciudadana dado la muerte de una menor que falleció a causa de un conductor irresponsable bajo los efectos del alcohol. En ella se estableció que se sanciona con cárcel efectiva de al menos un año a los conductores en estado de ebriedad que generen lesiones graves gravísimas o la muerte. Además, con esta reforma se establece como delito fugarse del lugar del accidente y negarse a realizar el alcohotest o la alcoholemia.

3.2. Los daños relacionados al alcohol

Los daños relacionados al alcohol están determinados por tres dimensiones del consumo de alcohol: el volumen de alcohol ingerido, el patrón de consumo y la calidad del alcohol consumido.

El consumo de alcohol se ha identificado como un elemento presente en la causalidad de más de 200 enfermedades, lesiones y condiciones de salud enfermedades, y otras condiciones de salud con códigos CIE-10.

3.2.1. Trastornos por consumo de alcohol

Existen diferentes dimensiones de consumo de alcohol que van desde el consumo de bajo riesgo hasta la dependencia a esta sustancia. A continuación se muestra una serie de definiciones para categorizar el consumo de alcohol.

Uso nocivo de alcohol (equivalente a consumo de riesgo de alcohol): se define como un patrón de uso del alcohol que provoca daños a la salud si el hábito de consumo persiste. En este caso el daño puede ser físico (por ejemplo, cirrosis hepática) o a nivel mental (casos de episodios depresivos posteriores a la ingesta de excesiva de alcohol) [52]. La OMS lo describe como el consumo regular de 20 a 40 gramos diarios de alcohol en mujeres y de 40 a 60 gramos diarios en varones.

Consumo perjudicial: se refiere a aquel que conlleva consecuencias tanto para la salud física como para la salud mental de la persona y está definido por la OMS como consumo regular promedio de más de 40 gramos de alcohol al día en mujeres y de más de 60 gramos de alcohol al día en varones. [41]

Consumo excesivo episódico o circunstancial: (en inglés conocido como *binge drink*) Implica el consumo, por parte de un adulto, de por lo menos 60 gramos de alcohol en una sola ocasión. Esto puede resultar particularmente dañino para ciertos problemas de salud.

Dependencia al alcohol: (también es conocido como alcoholismo o síndrome de dependencia del alcohol) se define como un grupo de fenómenos conductuales, cognitivos y fisiológicos que se desarrollan después del uso repetitivo de alcohol. Por lo general, se presenta un fuerte deseo de consumir alcohol, dificultades en el control de su uso, persistencia de su uso a pesar de conocer las consecuencias perjudiciales, dar mayor prioridad al consumo de alcohol que a otras actividades y obligaciones, aumento de la tolerancia, y a veces un estado de abstinencia

fisiológica [52].

3.3. Alcohol Use Identification Test (AUDIT)

El Test de Identificación de los Trastornos Debidos al Consumo de Alcohol, conocido en inglés como *Alcohol Use Identificación Test*, es un instrumento desarrollado por la Organización Mundial de la Salud (OMS) y que ha sido utilizado en Chile como tamizaje[2] para identificar a personas con consumo excesivo de alcohol, riesgo de dependencia y como una herramienta de apoyo en la intervención breve [7]. Este test es es un indicador que sugiere la existencia de consumo de riesgo de alcohol.

El AUDIT es un cuestionario de diez ítem que se separa en tres dominios, presentados en la tabla 3.1. La escala del Test tiene un rango entre cero y cuarenta puntos y la categoría de consumo de alcohol en la que se encuentra cada individuo va a estar dada por el puntaje total obtenido en el instrumento.

De acuerdo a las recomendaciones de la Organización Mundial de la Salud, los puntajes de corte que permiten identificar los distintos niveles de de riesgo en el test AUDIT son:

- Consumo de riesgo: entre 8 y 15 puntos.
- Consumo perjudicial: entre 16 y 19 puntos.
- Dependencia: 20 Puntos o más.

A pesar que la OMS sugiere estos valores de corte para cada nivel de riesgo, se sugiere que éstos deben ser seleccionados considerando los estándares culturales y nacionales [7], a través del desarrollo de estudios específicos para cada país.

Tal como se muestra en [4], la versión chilena del AUDIT fue validada el año 2009, en donde se incorporó la traducción y adaptación lingüística del instrumento. En el estudio [4] se concluyó que el instrumento es válido y confiable para el tamizaje de las distintas categorías de riesgo de consumo de alcohol en la población adulta chilena. En base a la sensibilidad y especificidad para cada categoría diagnóstica, se establecieron los puntos de corte en la escala los cuales son presentados a continuación:

- Consumo de riesgo: entre 6 y 8 puntos
- Consumo perjudicial o dependencia: 9 puntos o más

[2] pesquisa o screening

Dominio	Contenido del ítem
Uso peligroso de alcohol	Frecuencia de consumo Cantidad típica ingerida Frecuencia de consumo excesivo
Síntomas de dependencia	Deterioro del control sobre el consumo Aumento la importancia de beber que impide realizar las actividades cotidianas Consumo de alcohol durante las horas de la mañana
Consumo perjudicial de alcohol	Sentirse culpable después de beber Perdida de Memoria producto del alcohol Lesiones relacionadas al alcohol Preocupación de otros respecto al consumo de alcohol

Tabla 3.1: Dominios y contenidos de los items del Test AUDIT

Fuente: Elaboración propia a partir de [7]

Capítulo 4

Diseño

4.1. Requerimientos

En este capitulo se exponen los requerimientos del cliente, en este caso, el psiquiatra Carlos Ibáñez.

4.1.1. Variables originales de la Encuesta Nacional de Drogas

Las variables que tienen relación con la descripción del consumo y abuso de alcohol medidas en el Estudio Nacional de Drogas [65] para Población General se muestran a continuación:

1. Prevalencia de consumo
2. Percepción de riesgo
3. Disponibilidad de alcohol
4. Frecuencia de consumo
5. Embriaguez (*Binge Drinking*)
6. Consumo de riesgo de alcohol
7. Conducción de vehículos y consumo de sustancias
8. Consumo de sustancias en el hogar
9. Venta clandestina de alcohol
10. Aprobación de políticas de control de alcohol
11. Percepción de problemas de barrio de residencia

A continuación, se detalla lo que significa cada una de estas variables.

Prevalencia de consumo: corresponde a la proporción de individuos que ha consumido una determinada sustancia en un periodo de tiempo. Esta variable se usa para conocer el porcentaje de la población que usa la sustancia, pero no permite establecer el patrón de consumo. En el estudio se consideran tres ventanas temporales:

- **Durante la vida** o **Prevalencia de vida o global:** corresponde a la proporción de personas que consumieron la sustancia alguna vez en la vida sobre el total de las personas estudiadas.
- **El último año:** aquellas personas que declaran haber consumido al menos una vez alcohol en los últimos doce meses.
- **Último mes:** es la proporción de individuos que ha consumido alcohol al menos una vez en los últimos 30 días.

Percepción de riesgo de uso de alcohol: Corresponde al porcentaje de la población que percibe como una gran riesgo el consumo de cinco o más tragos de alcohol al día.

Oferta y disponibilidad de alcohol: en este indicador se pesquisa los lugares más frecuentes de compra de alcohol entre quienes han consumido en el último año. Para ello se le presenta al encuestado un listado de lugares establecidos y otro de lugares no establecido sobre los cuales debe declarar si realizó alguna compra durante los últimos treinta días. La lista de lugares establecidos y no establecidos se presenta a continuación.

Locales establecidos:

- Bares, pubs o discotecas.
- Restaurantes.
- Botillerías/licorerías.
- Supermercados.
- Servicentros.

Locales no establecidos:

- Ferias libres o mercados.
- Fiestas particulares.
- Locales clandestinos.
- A la entrada de conciertos, estadios, eventos.
- Productores artesanales.

Frecuencia de consumo: Para medir la intensidad de uso se emplea el número de días que en que se ha consumido alcohol en los últimos 30 días, calculado sobre aquellos individuos que reportaron consumo en el último mes.

Embriaguez (*Binge Drinking*): Embriaguez o *Binge Drinking*, como es conocido su nombre en inglés, corresponde a la proporción de individuos que declaró haber consumido cinco o más tragos en una sola ocasión (cuatro o más para el caso de mujeres) durante los últimos 30 días.

Consumo de riesgo de alcohol: Desde el año 2008, el SENDA utiliza un instrumento que permite identificar el consumo de riesgo de alcohol, además de otros transtornos asociados a su consumo. Para esto utiliza el puntaje del Test de Identificación de Transtornos Debido al Consumo de Alcohol (AUDIT), el cual como ya se ha señalado anteriormente tiene diez ítem que derivan en un rango de 0 a 40 puntos, donde ocho o más puntos se identifica con el consumo de riesgo. Para construir el indicador de consumo de riesgo de alcohol se conside-

raron a los consumidores del último año, que respondieron el instrumento AUDIT y además obtuvieron un puntaje superior a 8 puntos.

Conducción de vehículos y consumo de sustancias: Este indicador nace de la preocupación del desarrollo de políticas que buscan regular el uso de alcohol y otras drogas. Para esta sección se considera a aquellas personas mayores de 17 años que declaran poseer licencia de conducir. El objetivo es medir el porcentaje de individuos que condujo bajo la influencia del alcohol alguna vez en el último año.

Consumo de sustancias en el hogar: Existen investigaciones que señalan al hogar como factor protector o de riesgo para el consumo de sustancias, principalmente en la población joven y adolescentes. El consumo de sustancias en el hogar está asociado a una mayor disponibilidad y a una disminución de la percepción de riesgo frente al uso de drogas. El consumo de sustancias en el hogar se entiende como la proporción de individuos que declara que algún miembro del hogar consume alguna droga en particular. En este indicador también se estudia si a lo menos un miembro del hogar tiene problemas con el consumo de alcohol.

Aprobación de políticas de control de alcohol: proporción de individuos que se declara muy de acuerdo o de acuerdo con respecto a medidas para el control de alcohol.

Las medidas analizadas se muestran a continuación:

- Aumentar el impuesto a las bebidas alcohólicas.
- Reducir el número de locales que venden bebidas alcohólicas.
- Aumentar las penas para quienes conducen vehículos en estado de ebriedad.
- Reducir la hora límite para vender alcohol en las noches.

Percepción de problemas de barrio de residencia: en esta sección se presenta un conjunto de reportes asociados a fenómenos que ocurren en los barrios de las personas que fueron entrevistadas. El objetivo de esta caracterización es medir la percepción de ocurrencia de estos fenómenos y monitorear su tendencia a lo largo del tiempo.

Este indicador contiene ocho categorías, las cuales se enumeran a continuación:

1. Tráfico de drogas.
2. Robo en las casas.
3. Rayado en las paredes, daños al alumbrado o acciones similares.
4. Consumo de drogas en lugares públicos, como en la calle o plazas.
5. Asaltos o robos en las calles.
6. Jóvenes parados no haciendo nada en las esquinas.
7. Balaceras y acciones violentas con armas de fuego.
8. Venta clandestina de alcohol.

4.1.2. Segmentación original de la Encuesta Nacional de Drogas

En la encuesta se realiza una segmentación para poder estudiar de mejor manera a la población. Esta segmentación es realizada en en base a variables demográficas y geográficas. La encuesta toma cuatro variables de segmentación las que son detalladas a continuación:

1. Sexo: las categorías son:
   - Hombre
   - Mujer
2. Edad: la encuesta considera cinco rangos etarios:
   - 12 - 18
   - 19 - 25
   - 26 - 34
   - 35 - 44
   - 45 - 64
3. Nivel socio-económico: las personas se dividen en tres niveles dependiendo del nivel de ingresos declarado:
   - Bajo
   - Medio
   - Alto
4. Regiones: se subdivide en quince categorías, en la que cada una de ellas corresponde a una de las quince regiones administrativas de Chile.

4.2. Indicadores finales utilizados para el estudio en *Twitter*

La información en *Twitter* se presenta de forma diferente a como es preguntada en la Encuesta Nacional de Drogas. Por esta razón, es necesario realizar ciertas adaptaciones que permitan obtener la información necesaria para pesquisar el consumo de alcohol en esta plataforma.

Es importante aclarar que algunas de las segmentaciones realizadas en la encuesta no fueron posibles de replicar en este estudio en *Twitter*. Una de ellas es la segmentación por regiones, ya que no todos los usuarios explicitan la región en la que se encuentran. Además, agregar esta segmentación se traduciría en mayores condiciones para generar la red de usuarios chilenos, lo que escapa del alcance de esta memoria.

Es necesario destacar que los indicadores finales podrían diferir en un menor o mayor grado con la variable original explicada en la sección anterior 4.1.1. Sin embargo, estos aportan información al cliente para comprender el fenómeno del consumo de alcohol y la posterior toma de decisiones, tanto en políticas relacionadas al control del consumo de alcohol y su venta, como a nuevas opciones para incentivar a la población a un consumo responsable del alcohol.

Polaridad de *Tweets*: La clasificación de la polaridad es una tarea que categoriza un fragmento de texto, en este caso un *tweet*, en positivo, negativo o neutro dependiendo de su significado

emocional. El objetivo es poder identificar si la opinión expresada en relación al consumo de alcohol es negativa o positiva.

Prevalencia de consumo de alcohol: Este indicador busca pesquisar el porcentaje de individuos que consume alcohol. Será construido a partir de la información disponible en la cuenta del usuario, además del contenido generado por él.

Porcentajes de *following* o amigos consumidores de alcohol: Existen estudios en donde se pone en evidencia que los amigos ejercen una influencia trascendente en los hábitos de consumo de alcohol en los jóvenes. Como se muestra en [64] la influencia de los compañeros es considerada una de las principales causas de iniciación y persistencia del consumo de alcohol entre los adolescentes. Esto no es tan sorpresivo, ya que al igual que el caso de los adultos, beber alcohol y emborracharse son principalmente actividades para sociabilizar con sus pares. Con este indicador se busca descubrir si en el caso de las redes sociales se produciría un efecto similar, osea identificar si una persona al estar expuesta a lo que sus amigos comparten en relación al alcohol, se vería influenciada por ellos o presentaría patrones de consumo de alcohol similares.

Palabras utilizadas: En este ítem se pretende obtener información acerca de las preferencias de bebidas alcohólicas por parte de los usuarios, ya sea por las bebidas alcohólicas destiladas o las fermentadas, y obtener información acerca de la calidad del alcohol consumido. Según [61], a la hora de analizar las consecuencias generadas por el consumo, una de las variables relevantes es la calidad del alcohol consumido. También, se analizará si el vocabulario utilizado varía según el contexto, es decir, se identificará si existe diferencias en cuanto a las palabras claves utilizadas cuando se refiere a las políticas de alcohol contrastado al utilizado para comentar sobre el consumo de alcohol.

Polaridad de *tweets* de políticas: En este indicador se busca conocer las opiniones acerca de las políticas implementadas o ideas de políticas a futuro relacionadas con el control de consumo y venta de alcohol. La idea es poder tener mayor información sobre la reacción de las personas a nuevas políticas.

Frecuencia de consumo: En este indicador se medirá la frecuencia de *tweets* generados por los usuarios que implica explícitamente el consumo de alcohol o también cuando se puede deducir el consumo de manera indirecta.

4.3. Descripción de los datos utilizados

En esta sección se describe el tipo de información que se utiliza como dato de entrada. Primero, se habla de la información extraída utilizando la API de *Twitter*. A continuación, se detalla la estructura que deben tener los datos para el entrenamiento de los algoritmos de aprendizaje. Finalmente, se detallan los mecanismos utilizados para realizar el etiquetado de tweets y de usuarios para realizar los análisis y algoritmos de clasificación.

4.3.1. Datos disponibles

Para la extracción de información se utilizó la red social Twitter, ya que el espíritu de esta red social es que la cuenta sea pública. Sin embargo, existen cuentas privadas a las cuales no se puede acceder a la información a no ser que el propietario de esa cuenta acepte la solicitud de seguimiento. A pesar de eso, el número de cuentas privadas es menor en relación a las cuentas públicas.

Además, se cuenta con datos históricos recolectados en la base de datos *La Gorda* disponible en el Centro WIC. Esta base de datos es constantemente actualizada por el equipo de *Opinion Zoom* y tiene la ventaja de tener más datos disponibles para el desarrollo de los algoritmos.

La información útil para el presente proyecto disponible en *Twitter* es la siguiente:

- Información acerca del usuario.
- Red del usuario conformada por sus conexiones con otros usuarios, es decir, las relaciones de seguimiento.
- *Tweets* publicados por el usuario.

Los datos necesarios para llevar a cabo la investigación serán obtenidos a través de la API REST de *Twitter*. Para hacer uso de ella se necesita conocer el identificador único que posee cada usuario, el cual puede ser el número de identificación o el nombre de usuario. Los *tweets* también poseen un número de identificación único, lo que ayuda a la tarea búsqueda de un *tweet* en particular. Los datos son entregados en formato JSON. En la tabla 4.1 se observa los datos del usuario utilizados para el presente trabajo. Cabe hacer notar que es posible rescatar más información del usuario, pero en este caso eso no es relevante. Por otro lado es importante destacar que el nombre del usuario es elegido por el propio usuario, que podría o no corresponder a su nombre real, de todas forma para efectos de este estudio este campo no fue utilizado, por lo que no se vió comprometida la privacidad de las personas

En la tabla 4.2 se observa la información del *tweet* que será utilizada. Al igual que en el caso anterior, la tabla sólo rescata los campos que serán de ayuda para desarrollar posteriormente los modelos.

4.3.2. Estructura de los datos

Para poder hacer uso de la información disponible en Twitter, es necesario realizar ciertas transformaciones y un procesamiento previo a los datos. Se cuenta con datos estructurados y no estructurados. Un ejemplo de campos no estructurado son aquellos que contienen texto, los cuales por definición no cuentan con una estructura definida y para los cuales se necesita de preprocesamiento para posteriormente poder aplicar técnicas de reconocimiento de patrones.

Primero se necesita clasificar los *tweets* con respecto a tres categorías independientes.

- Consumo de alcohol (variable binaria):
  - 0 en el caso que no corresponda a consumo de alcohol.

Nombre	Descripción del dato	Tipo de Dato
ID	Número único identificador	Bigint
Screen Name	Nombre identificador	Texto
Name	Nombre del usuario	Texto
Description	Descripción que el usuario hace de si mismo	Texto
Lang	Lenguaje	Texto
Location	Ubicación	Texto
Timezone	Zona horaria	Texto
Createdat	Fecha y hora de la creación de la cuenta	Tiempo
Followerscount	Número de seguidores	Entero
Friendscount	Número de amigos o personas a quienes sigue el usuario	Entero
Statusescount	Número de tweets	Entero
Isgeoenabled	Opción de geo-localización	Booleano
Isprotected	Si los *tweets* se encuentran protegidos, es decir corresponde a una cuenta privada	Booleano

Tabla 4.1: Datos del usuario disponibles en *Twitter* y útiles para el estudio.

Fuente: Elaboración Propia

Nombre	Descripción del dato	Tipo de Dato
ID	número único identificador	Bigint
User	Usuario asociado a la autoría del tweet en cuestión	Bigint
Text	Contenido del *tweet*	Texto
Istruncated	Si es tweet corresponde a un mensaje cortado	Booleano
Createdat	Fecha y hora de creación del n*tweet*	tiempo

Tabla 4.2: Datos del *tweet* disponibles en *Twitter* y útiles para el estudio.

Fuente: Elaboración Propia

  - 1 en caso que del tweet se pueda deducir consumo de alcohol.
- Mención de políticas de control de alcohol (variable binaria):
  - 0 en el caso que en el *tweet* no se mencionen políticas de alcohol.
  - 1 en el caso que el tweet haga alusión a alguna política de alcohol.
- Venta de alcohol (variable binaria):
  - 0 en el caso que no se pueda inferir venta de alcohol por parte del autor del *tweet*.
  - 1 en el caso que se pueda inferir explicita o implícitamente la venta de alcohol por parte del autor del *tweet*.

En segundo lugar se utilizará la información disponible como *tweets*, información del usuario y conexiones entre la red de usuarios para determinar dos atributos en los usuarios.

- Consumo de alcohol (Variable binaria)
  - 0 en el caso que el usuario no haya consumido alcohol el último año
  - 1 en el caso que se deduzca que el usuario ha consumido alcohol el último año
- Rango de edad (variable categórica)

Para poder entrenar los algoritmos de aprendizaje supervisado, es necesario tener un conjunto de datos etiquetados y además tener la certeza que estos datos estén etiquetados correctamente. Este proceso de etiquetado será descrito en las secciones de Etiquetado de *Tweets* 4.3.3 y Etiquetado de usuarios 4.3.4

4.3.3. Etiquetado de *Tweets*

Para un ser humano, el extraer información relevante de un texto no es una tarea extremadamente difícil. El problema empieza cuando se necesita clasificar una cantidad enorme de información, lo que requeriría de muchas horas de trabajo, que a largo plazo no resulta rentable. Debido a esto surge la necesidad de integrar mecanismos automáticos que puedan realizar esta tarea.

4.3.4. Etiquetado de usuarios

En esta sección, los usuarios se asignan a si mismos las etiquetas. Este paso es necesario para poder hacer el estudio acerca de la prevalencia de alcohol. Se necesita que el usuario entregue información sobre su consumo de alcohol, y algunas situaciones que están asociadas al abuso de esta sustancia. Es importante destacar que este indicador está acotado al contexto de las redes sociales, por lo que hay que tener presentes las limitaciones propias de estas.

El etiquetado será realizado por medio de una encuesta directa enviada a los usuarios. Primero, se construirá una base de datos con los usuarios chilenos de *Twitter*. A continuación, se toman usuarios al azar para enviarles la encuesta. La forma de difundir la encuesta será primero creando una cuenta de *Twitter* destinada a la difusión de la encuesta, y a continuación publicando un *tweet* mencionando al usuario (@usuario) para que de esta forma el usuario reciba una notificación. La razón de elegir esta forma de difusión es que a pesar de que en *Twitter* existen los menajes

directos, para que esta opción esté disponible se necesita que el usuario al cual se desea enviar el mensaje, previamente sea un seguidor de uno, por lo que para fines prácticos la alternativa detallada anteriormente es la más apropiada.

La encuesta contendrá preguntas que entreguen información acerca del consumo de alcohol, la edad, el sexo y el consumo de riesgo y dependencia de cada usuario. Finalmente, el usuario deberá ingresar el nombre de usuario de su cuenta de *Twitter* para así poder relacionar los resultados de la encuesta con los datos recogidos de *Twitter*.

4.3.5. Selección de *Keywords*

Para clasificar y poder seleccionar los *tweets* relacionados con alcohol, se necesita contar con palabras claves o *Keywords*. La idea es filtrar de la gran masa de *tweets* existentes en esta red social, solamente los *tweets* que tienen relación con el alcohol.

La lista de palabras claves tendrá origen en una encuesta que fue realizada por cuenta propia. La razón de utilizar una encuesta es poder obtener los términos actuales que se utilizan para referirse al alcohol, ya que estos términos varían a lo largo del tiempo. La encuesta contendrá solamente una pregunta: "¿Cuál o cuáles son los términos o expresiones que usted utiliza, ha escuchado o conoce para referirse al alcohol o a su consumo?". Acá el encuestado podrá escribir todos los términos que desee.

Posteriormente serán ordenadas de manera descendente y se seleccionarán las palabras más repetidas.

A continuación se verificará el uso de estas palabras claves en *Twitter*, con el objetivo de desambiguar el contexto de uso y siendo consciente que el lenguaje utilizado por los usuarios en *Twitter* podría variar del lenguaje que es utilizado en otros ambientes. Para hacer el proceso de desambiguación desde *La Gorda* se extraerán *tweets* que contengan las palabras.

Se aplicará un preprocesamiento de texto el cual consta de:

- Tokenización
- Remoción de *stopwords* o palabras vacías
- Eliminación de emoticones y *emojis* (por ejemplo ":)")
- Eliminación de elementos propios de *Twitter* por ejemplo, *Retweets*

Posteriormente, será aplicado *Topic Modeling* con el propósito de identificar distintos contextos de uso que se le da a esas palabras en *Twitter* y verificar que realmente estén relacionadas con el uso y consumo de alcohol. Para cada una de las palabras, lo que hará *Topic Modeling* es agrupar los *tweets* en distintos grupos, los cuales serán determinados por este modelo y manualmente se deberá identificar el *clúster* que habla de alcohol y las palabras que lo caracterizan.

En el caso que existan palabras ambiguas, estas serán filtradas considerando solamente aquellos *tweets* que además de la palabra analizada contengan la cadena de caracteres "tom", el que hace alusión al verbo "tomar". Por lo anterior es que solo se dejan las tres primeras letras, para así poder

considerar también todas las conjugaciones que existen de este verbo (tomo, tomé, tomábamos, etc), que de un análisis exploratorio se concluyó que es la palabra más comúnmente utilizada en este contexto.

Se consideró una segunda regla, además de la anteriormente señalada "tom", para filtrar *tweets* relacionados con alcohol. Este es el caso del uso de la herramienta *Freeling*. Esta regla solamente fue utilizada para la palabra vino, debido a la importancia que tiene en nuestro país esta bebida alcohólica, no sería aconsejable eliminarla de la lista de palabras claves. Sin embargo, al incluir el desambiguador "tom" resultaba ser demasiado restrictivo. Por lo tanto se utilizó el hecho que vino puede cumplir dos funciones a nivel gramatical, puede ser un verbo, vino de venir, o también puede representar a un sustantivo, la bebida alcohólica vino.

Es posible utilizar la herramienta de *Freeling* de dos formas distintas:

- Obteniendo el lema de la palabra, para este trabajo el caso de interés es cuando el lema es igual a vino. El caso que se debía descartar es cuando el lema es igual a ir.
- Obteniendo el *tag* de la palabra. Pueden ocurrir dos posibles casos para la palabra vino, que el tag sea igual a n (*noun* o sustantivo) y v (*verb* o verb). En el caso particular de este estudio se necesitaba que el *tag* fuera igual a *noun*.

De las dos posibles formas de abordar el problema se escogió la primera, pero ambas son equivalentes y se obtendría resultados similares.

Esta segunda regla solamente fue aplicada a la palabra vino, debido a los dos usos distintos que puede tener a nivel gramatical. Existen otras palabras que son ambiguas, pero que tanto para el caso que es de interés y el que se desea descartar la palabra cumple la función de sustantivo. Un ejemplo de esto es la palabra terremoto, que a pesar de utilizar el lema o el tag es muy difícil desambiguar de manera computacional.

Es importante destacar que el uso de esta segunda regla se traduce en que se requiere de mayor tiempo de procesamiento para hacer el filtrado.

4.4. Diseño de la aplicación

4.4.1. Tratamiento de texto

Para poder trabajar con los **tweets** fue necesario realizar una proceso de pre-procesamiento el cual consistió en las etapas que se enumeran a continuación:

1. Tokenización
2. Normalización
3. Eliminación de caracteres especiales
4. Eliminación de elementos de *Twitter*, tales como @mencion o RT
5. *Stemming*

6. Formulación de n-gramas
7. Remoción de Palabras Vacía, en inglés conocidad como *Stopwords*
8. Representación de documentos (vectores)

La clasificación de texto es del tipo binario, es decir, se tienen dos categorías en la que un texto puedo pertenecer o no a esa clase. Las categorías a evaluar son:

- consumo de alcohol por parte del autor del texto.
- mención de políticas relacionadas con alcohol.
- venta de alcohol por parte del usuario.

Cada una de las categorías anteriormente mencionadas será evaluada de forma independiente de las otras. Estas categorías no se superponen y por lo tanto un *tweet* puede pertenecer a una, dos o todas las categorías descritas.

Para cada una de las categorías será construido un clasificador distinto.

Para construir los clasificadores se utilizarán algoritmos, que según la literatura, han demostrado tener un mejor despeño en cuento a clasificación de texto se refiere [28]. Estos algoritmos son:

- *Support Vector Machines*
- *Voted Perceptron*
- *Naive Bayes*
- Árboles de Decisión

4.4.2. Cálculo de la polaridad de *tweets*

Para realizar el cálculo de la polaridad de los *tweets* se utilizará una API de *Opinion Mining*, perteneciente al proyecto OpinionZoom desarrollada en el centro WIC, y cuyo nombre es PAPI.

4.4.3. Cálculo de la edad en usuarios

Para realizar la predicción de la edad de los usuarios se utilizará un algoritmo desarrollado anteriormente en el centro WIC, el cual hace una estimación de la edad de un usuario en particular, a partir del análisis de los últimos doscientos *tweets* publicados por el usuario.

4.4.4. Atributos para el usuario

El conjunto de atributos utilizados para clasificar el consumo de alcohol en usuarios es el siguiente:

- Número de menciones de alcohol

- Número de menciones de consumo
- Número de menciones de políticas
- Métricas de polaridad del usuario
  - Polaridad de políticas
  - Polaridad de los *tweets* de consumo
- Polaridad de vecindario
- Edad
- Métricas de Análisis de Redes Sociales (ARS)
  - Métricas de incrustación (*embeddedness*)
    - Densidad de Vecindario
    - Nominaciones externas
  - Métricas de centralidad (*social status*)
    - *Normed indegree*
    - *Reach centrality* (2 nominaciones)
  - Proximidad a usuarios que hacen mención de consumo
    - Porcentaje de consumidores en vecindario
    - Distancia mínima al algún consumidor
- Ponderación temporal de menciones de alcohol
- Ponderación temporal de menciones de consumo
- Ponderación temporal de menciones de políticas

El conjunto de atributos utilizados para clasificar el consumo de riesgo y dependencia en usuarios es el siguiente:

- Consumo de alcohol
- Número de menciones de alcohol
- Número de menciones de consumo
- Número de menciones de políticas
- Métricas de polaridad del usuario
  - Polaridad de políticas
  - Polaridad de los *tweets* de consumo
- Polaridad de vecindario
- Edad
- Métricas de Análisis de Redes Sociales (ARS)
  - Métricas de incrustación (*embeddedness*)
    - Densidad de Vecindario
    - Nominaciones externas
  - Métricas de centralidad (*social status*)
    - *Normed indegree*
    - *Reach centrality* (2 nominaciones)

    - Proximidad a usuarios que hacen mención de consumo
        * Porcentaje de consumidores en vecindario
        * Distancia mínima al algún consumidor
- Ponderación temporal de menciones de alcohol
- Ponderación temporal de menciones de consumo
- Ponderación temporal de menciones de políticas

Capítulo 5

Implementación

5.1. Herramientas utilizadas

En este apartado se describen las tecnologías escogidas que pueden hacer posible el sistema de información.

Ubuntu

Es un sistema operativo basado en GNU/Linux, el cual es distribuido como software libre. Este incluye su propio entorno de escritorio denominado Unity. Su nombre proviene de la ética homónima, en la que se habla de la existencia de uno mismo como cooperación de los demás. A continuación se detallan las ventajas de Ubuntu:

- Es gratuito.
- Uno de los puntos más importantes es la seguridad. Rara vez los creadores de virus tienen en su objetivo algún software de Linux.
- Permite cargar y realizar tareas con mayor eficiencia que otros sistemas operativos.
- Constantemente están apareciendo actualizaciones y nuevas versiones.

NetBeans

La versión utilizada de NetBeans es la 8.1[1]. Netbeans es un entorno de desarrollo integrado de código abierto. En él se puede trabajar con variados lenguajes de programación incluyendo Java, PHP, C/ C++ y Python, entre otros. Permite agilizar la programación leyendo el código del proyecto completo, lo cual en Java, es sumamente útil para evitar errores de compilación. Entre sus ventajas destaca que es una herramienta gratuita y existe una gran comunidad de usuarios y desarrolladores alrededor del mundo.

[1] https://netbeans.org/

Java

Este es un lenguaje de programación que cuenta con un conjunto de herramientas comerciales y de software abierto. Es uno de los lenguaje de programación más populares según el índice TIOBE, esto es de gran ayuda a la hora de buscar información en la red. Entre las caraterísticas se encuentra que este lenguaje no depende del *hardware* ni del sistema operativo y además está orientado a objetos lo cual es de interés para los desarrolladores.

FreeLing

Es una librería de código abierto de C++ que provee un servicio de análisis de lenguaje orientado a desarrolladores. La versión utilizada de Freeling[2] provee identificación del lenguaje, tokenización o segmentación de palabras, división de frases, análisis morfológico, codificación fonética, etiquetado POS, análisis sintáctico superficial[3]. El proyecto FreeLing tuvo sus inicios en el Center for Language and Speech Technologies and Applications (TALP)[4] de la Technical University of Catalunya (UPC), con el objetivo de avanzar hacia la disponibilidad general de recursos y herramientas básicos de Procesamiento del Lenguaje Natural (PLN). Esta disponibilidad debería posibilitar avances a mayor velocidad en proyectos de investigación y poder reducir los costos en el desarrollo de aplicaciones de PLN.

PostgreSQL

Es un sistema de gestión de bases de datos relacional orientado a objetos. Es *Open Source* o también llamado código abierto.

Algunas de las ventajas de PostgreSQL por sobre otras plataformas se detallan a continuación [27]:

1. Open Source: está desarrollado con código fuente abierto u *Open Source*. En su pagina web podemos encontrar el código y descargarlo gratuitamente, lo que entrega libertad de desarrollo que permite a colaboradores de alrededor del mundo incluir nuevas mejoras. El desarrollo de PostgreSQL no es manejado por una empresa o persona, sino que es dirigido por una comunidad de desarrolladores que trabajan de forma desinteresada, altruista y libre. Dicha comunidad es denominada *PostgreSQL Global Development Group* (PGDG).
2. Multiplataforma: se encuentra disponible para una amplio número de versiones de los sistemas operativos de Unix y Windows, lo que ayuda a que no existan problemas de compatibilidad.
3. Alto volumen: Para bases de datos de gran volumen PostgreSQL tiene un gran rendimiento. Gracias al Método de Control de Concurrencias Multiversión, o también conocido por su nombre en inglés *Multiversion Concurrency Control*, cuya sigla es MVCC, ayuda a tener una

[2] http://nlp.lsi.upc.edu/freeling/
[3] en inglés llamado shallow parsing
[4] http://www.talp.upc.edu/

mejor *performance* cuando hay muchos movimientos en la base datos. El objetivo principal de este método es permitir leer y escribir de forma simultánea, es decir, sin que ninguna de las dos operaciones bloquee a la otra.

4. Facilidad de manejo: PgAdmin es un entorno de escritorio visual para la gestión y administración de base de datos PostgreSQL. Esta herramienta está pensada para facilitar el trabajo a cualquier usuario, desde el principiante hasta el más avanzado. Entre sus características destaca que permite acceder a todas las funcionalidades de la base de datos.
5. Seguridad de la información: *Hot Standby* es el nombre de la capacidad de poder ejecutar consultas a una base de datos que simultáneamente está realizando una recuperación de archivos. La replicación de envío de registros permite crear uno o más nodos de reserva que son réplicas del nodo principal o el también llamado nodo maestro. Los nodos en espera pueden ser utilizados para realizar consultas solamente de lectura[60]. En otras palabras, la incorporación de *Hot Standby* permite a los usuarios acceder a las tablas en modo lectura mientras se realizan los procesos de backup o mantenimiento, sin que se vea comprometida la integridad de los datos.

Estas características hacen de PostgreSQL una de las bases de datos más avanzada y potente del mercado.

JSON

La sigla es un acrónimo de *JavaScript Object Notation*. Este es un formato de texto ligero para el intercambio de datos, además de ser autodescriptivo y fácil de entender. Esto significa que para humanos leerlo y escribirlo es simple, mientras que para las máquinas es simple interpretarlo y generarlo.

JSON es un formato de texto completamente independiente del tipo de lenguaje, lo que quiere decir que el texto puede ser leído y utilizado como un formato de datos por cualquier lenguaje de programación. A pesar de eso, utiliza convenciones que son ampliamente conocidos para los programadores de la familia de lenguaje C, incluyendo C, C++, C#, Java, JavaScript, Perl, Python, entre otros. Dado que formato JSON es solamente texto, puedo ser fácilmente enviado desde un servidor y ser utilizado como formato de datos por cualquier lenguaje de programación. Las características nombradas anteriormente hacen de JSON un lenguaje ideal para el intercambio de datos.

En la figura 5.1 se muestra un ejemplo del formato JSON.

Twitter4J

Es una librería no oficial para trabajar con la API de Twitter. Esta librería permite integrar fácilmente las aplicaciones desarrolladas en Java con los servicios de Twitter orientados a desarrolladores, como son la Stream API y la Rest API. Contiene recursos para el manejo de credenciales, necesarias para acceder a la mayoría de datos contenidos en *Twitter*, como por ejemplo, información de usuarios, *tweets* y relaciones de seguimiento.

```
{
"responsable":
    {
    "Nombre" : "Juan",
    "Edad": 28,
    "Aficiones": ["Música", "Cine", "Tenis"],
    "Residencia": "Madrid"
    },
"empleados":
    [
        {
        "Nombre" : "Elena",
        "Edad": 26,
        "Aficiones": ["Música", "Cine"],
        "Residencia": "Madrid"
        },
        {
        "Nombre" : "Luis",
        "Edad": 31,
        "Aficiones": ["Teatro", "Cine", "Fútbol"],
        "Residencia": "Madrid"
        }
    ]
}
```


Figura 5.1: Ejemplo de formato JSON.

Fuente: elaboración propia

JGibbLDA

JGibbLDA es una implementación en Java de Latent Dirichlet Allocation (LDA) utilizando la técnica de *Gibbs Sampling* para la estimación de parámetros y la inferencia. Este ejemplo de *Topic Modeling* permite obtener todos los parámetros calculados por LDA a partir de un archivo que contiene los textos a analizar. Los parámetros del modelo deben ser ingresados de forma manual.

JGibbLDA es útil para las siguientes potenciales áreas de aplicación:

- Recuperación de información (análisis semántico, temas latentes, gran colección de texto para una búsqueda de información inteligente).
- clasificación de documentos, *Clustering*, resumen de textos y para la comunidad de *Text/Web Data Mining* en general.
- Filtración colaborativa.
- Clusterización de imágenes basado en el contenido, reconocimiento de objetos y otras aplicaciones de *Computer Vision* en general.
- Otras aplicaciones potenciales destinado al estudio de datos biológicos.

JGibbLDA es un software libre y el usuario puede redistribuirlo y/o modificarlo bajo los términos de la liciencia GNU General Public License publicada por la *Free Software Foundation*.

Weka

Esta es una extensa colección de algoritmos de aprendizaje automático o *Machine learning* para realizar Minería de Datos, desarrollado por la Universidad de Waikato en Nueva Zelanda. El nombre Weka hace alusión a un ave endémica de Nueva Zelanda. Llamada científicamente como Gallirallus Australis, físicamente posee un color pardo y de tamaño similar a una gallin que no tiene la capacidad de volar, se encuentra en peligro de extinción. Es conocida por su curiosidad y agresividad. Weka es un software de código abierto cuya licencia es GNU General Public License, lo que significa que este programa es de libre uso, distribución y difusión.

Las razones de la utilización de Weka son:

- Está disponible libremente bajo la licencia pública general de GNU.
- Es portable, ya que está implementado en Java y se puede correr en numerosas plataformas.
- Contiene una extensa librería de técnicas destinadas al pre procesamiento de datos y modelamiento.

Ficheros .arrf

Nativamente Weka funciona con un formato llamado *arff*. Este es un acrónimo de *Attribute-Relation File Format*. Un archivo arff es un archivo de texto ASCII que describe una lista de instancias que comparten un conjunto de atributos. Este formato fue desarrollado por el Proyecto de *Aprendizaje de máquinas* del Departamento de Ciencias de la Computación de la Universidad de Waikato con el fin de ser usado con el software Weka. Los archivos ARFF tienen tres secciones distintas. La primera sección se conoce como *header* que corresponde al encabezado, luego continua la declaración de atributos, y finalmente sigue la información con la *Data*.

Encabezado. En esta parte se define el nombre de la relación. La sintaxis es:

$$@relation \ < nombre > \tag{5.1}$$

El tipo de dato soportado por <nombre> es de tipo *String* y en el caso de contener espacios debe escribir entre comillas.

Declaración de atributos. En esta sección se declaran los atributos que componen nuestro archivo junto a su tipo. Los atributos corresponden a las columnas de la data. La sintaxis es la siguiente:

$$@attribute < nombre\_del\_atributo > \ < tipo\_de\_dato > \tag{5.2}$$

En donde <nombre_del_atributo> debe ser de tipo *String*. Para <tipo_de_dato> Weka acepta los siguientes:

1. Numeric: Para números reales.

2. Integer: Para expresar números enteros.
3. Date: Destinada a expresar fechas, para ello este tipo de dato debe ir precedido de una etiqueta de formato entrecomillas. La etiqueta de formato está compuesta por caracteres separadores (guiones y/o espacios) y unidades de tiempo:
   - dd Día.
   - MM Mes.
   - yyyy Año.
   - HH Horas.
   - mm Minutos
   - ss Segundos.
4. String: para expresar cadenas de texto.
5. Enumerado: este tipo consiste en escribir entre llaves y usando como separación las comas los posibles valores (caracteres o cadena de caracteres) que puede tomar el atributo. A continuación se muestra un ejemplo considerando que se tiene un atributo que indica el tiempo.

$$@attribute\ tiempo\ \{soleado, lluvioso, nublado\} \tag{5.3}$$

Sección de datos En esta sección se deberá declarar los datos que componen la relación separando entre comas los atributos y con saltos de línea las relaciones. Todas las filas deberán tener el mismo número de columnas. Este número es el mismo que el de las declaraciones de atributos que fueron añadidas en la sección de declaración de atributos. En el caso que algún dato sea desconocido se deberá expresar con un símbolo de cierre de interrogación ("?"). Además, se pueden añadir comentarios con el símbolo de porcentaje ("%"), los que pueden situarse en cualquier lugar del fichero. La sintaxis es la siguiente:

@data
4,3,2

En la imagen 5.2 se muestra un archivo .arff a modo de ejemplo.

ArangoDB

ArangoDB es una base de datos NoSQL multi-modelo que fue desarrollada por triAGENS GmbH. Es una de las bases de datos NoSQL mayormente utilizada de las que actualmente se encuentran disponibles y que cuentan con una licencia de código abierto. Los creadores se refieren a ella como una base de datos multi-modelo nativo, ya que al usuario se le permite almacenar sus datos como pares llave/valor.

5.2. Selección de palabras claves o Keywords

El proceso de selección de palabras claves es un paso importante ya que dicta las pautas de la información que se buscará en Twitter.

```
@relation Rel

@attribute label {0,1}
@attribute text string

@data
0,'unas chelas varias < 3 @mention bar http'
0,'champanazo gratis ahora en el bazar de nomadesert ! hasta las 9pm http'
0,'su buen vinito ! ! !'
1,'#hashtag el refrigerador con cerveza lo llenaria , me tomo cada una hasta que la muerte me llame'
1,'ready pa la sed mojito tkm uwu #hashtag http'
0,'@mention y si iban a160 por haber tomado dos tragos ?'
1,'@mention @mention yo con vodka tonica .'
0,'i liked a @mention video from @mention http 42 frases tipicas de conductores ebrios'
0,'me tomaria un vodka ahora , pero manana tengo que aplicarme con el pedal antes de que salga el sol ! ! ...'
1,'@mention _ : son las 2 y estoy hecho concha // el peor combinado es el celular con internet + alcohol . te da cana seguro .'
0,'@mention el alcohol mata neuronas ( ministerio de salud )'
1,'@mention creo firmemente que estoy sobre ejercitado ! ! . jaja ! ! ! me estoy esforzando demasiado . ! ! ! mas me gusta el vino . . ! !
1,'viendo el partido de la tt_tt ( @mention coquimbo ) http'
0,'fiesta de la cerveza artesanal - valdivia 2014 : http de pelos ! ! !'
0,'#hashtag con 5 piscolas la encuentro rica'
1,'acto n1 : vaso en el frezeer ; acto n2 : juguito de limon ; acto n3 : cerveza heladita . como se llama la pelicula ?'
1,'@mention : buenos dias ! ! dr. hace un tiempo se me adormesen la punta de los dedos y me duele un poco el , , , // faltan chelas y entra
0,'@mention grande tomas ! ! ! ! ! ! orgullosa de ser chilena al ver tu orgullo al mostrar nuestro escudo patrio . manana sera un gran di
1,'insisto : no hay nada mejor que un sabado con pizzas , piscolas , casita , boxeo o vale todo . si o no , estimado @mention ?'
1,'quiero una michelada para que se me pase este puto resfriooo'
1,'para estos dias calurosos nada mejor que una rica y helada cerveza ! ! #hashtag'
1,'un sentido adios : ¡ salud por los @mention ! las pilsen de esta noche seran en su honor y la cana tambien http'
0,'este mino es alcoholico como consejo deja de creerte riiiico por ke no lo eres y ademas deberias hacerte un tratamiento de alcohol weeet
0,'@mention esta curao ese wn jajaja'
1,'necesito verano , calor , playa , cerveza , etc ! #hashtag'
0,'y dales las gracias por andar curao y por irresponsabilidad de todo esto @mention @mention'
0,'recuerden no consumir ni bombones rellenos de licor porque arrojan borrachera en el alcotest ; )'
1,'17:30 hrs en todo el territorio nacional continental oficialmente de vacaciones ! disponible solo para familia y chelas'
0,'@mention @mention oh si como dice la babo son como el vino e_e'
0,'without alcohol http'
```

Figura 5.2: Ejemplo de archivo Arff

Fuente: elaboración propia

Las fuentes de donde se extraen las keyword son:

- Una encuesta compartida a través de las redes sociales.
- Principales marcas de cervezas comercializadas en retailes del país (botillerías y supermercados).

Para obtener las palabras o keywords a buscar, se difundió una encuesta en donde la persona podía escribir libremente cualquier término que haya leído o escuchado y que asocie al consumo o efectos relacionados con el alcohol, por lo que no se descartó ningún término a priori. También, se incluyeron las principales marcas de bebidas fermentadas y destiladas disponibles en el mercado nacional, como botillerías y supermercados (por ejemplo, Báltica). Fueron incluidos los nombres genéricos de las bebidas alcohólicas (como cerveza, vino, pisco) y finalmente, algunas variaciones de las cepas de vino (merlot, carmenere)

Luego, se ordenaron los términos según la frecuencia en la que aparecieron en la encuesta. En este paso se descartaron los que aparecieron solo una vez, ya que dada la cantidad de respuestas que se obtuvo se consideró que si un término se repetía solo una vez, entonces no era muy usado.

Posteriormente, por cada uno de los términos seleccionados, se tomó una muestra de los tweets almacenados en la Gorda (máximo 5000 tweets) que contuvieran cada una de las palabras.

A continuación, se verificó que las palabras fueran utilizadas para referirse al alcohol en el contexto de Twitter. Esto debido a que las formas en que las personas se expresan por escrito en la redes sociales podría ser distinto a como se comunican oralmente. Este procedimiento se realiza para evitar tener términos ambiguos que afectaran al rendimiento de los modelos.

Para esto se utilizó Toping Modeling en el caso que la palabra tuviera más de 100 tweets, en caso contrario, es decir, si se contaba con una muestra menor a 100 tweets, se etiquetó manualmente cada uno de los tweets y se calculó el porcentaje de tweets que efectivamente estaban relacionados a alcohol y fueron conservados aquellos que tuvieran una tasa mayor a 60 %.

Para los términos que presentaron ser ambiguos, por ejemplo terremoto, que es utilizado para movimientos telúricos y para un popular brebaje, se utilizó la palabra tom.

5.3. Etiquetado

Para el presente trabajo se necesitó realizar el etiquetado de *tweets* y de usuarios de la red social. Para poder entregar un entorno más amigable con el usuario, fueron creados dos sitios Web, cada uno de ellos destinado a distinto tipo de etiquetado. Se utilizó la herramienta de desarrollo CodeIgniter, utilizando la arquitectura Modelo-Vista-Controlado (MVC). A medida que los usuarios iban avanzando, el sitio web iba realizando la escritura por medio del gestor de bases de datos relacionados PostgreSQL.

5.3.1. Etiquetado de *Tweets*

Primer etiquetado de *Tweets*

El etiquetado de *tweets* en las categorías seleccionadas se realizó mediante el uso de sesiones. A los etiquetadores se les entregó un usuario y clave con los cuales podía ingresar al sitio. Primero, se visualizaban las instrucciones para el etiquetado y también algunos ejemplos para mayor claridad del procedimiento que debían realizar posteriormente. Se incluyó además la lista completa con todas las palabras claves (*Keywords*) para que en el caso que el etiquetador tuviera alguna duda pudiera consultar fácilmente los términos que están asociados o hacen referencia a las bebidas alcohólicas, su consumo y su abuso. A continuación, se mostraban uno a uno los *tweets* que debía etiquetar.

El flujo de *tweets* era lineal: los casos a etiquetar iban apareciendo uno después del otro a medida que el usuario avanzaba en el proceso. Las etiquetas eran guardadas de manera automática, una a una a medida que el etiquetador hacía click en "siguiente *tweet*". Lo relevante de esto es que no era necesario que el etiquetador completara todo el proceso en una sola sesión, es decir, podía completar un número de *tweets*, cerrar la sesión, posteriormente volver a entrar y su sesión estaría tal como la dejó la última vez. Así, el etiquetador podía aprovechar pequeños momentos libres para continuar con el etiquetado.

También, el usuario podía revisar los casos anteriormente clasificados por él mismo y deshacer el trabajo realizado. En el caso que el usuario deshiciera algún *tweet* etiquetado, éste *tweet* sería puesto al final de la cola de *tweets* a etiquetar.

Luego, se mostraba un mensaje indicando que la lista de *tweets* disponibles para etiquetar para

ese usuario en particular habían llegado a su fin. Finalmente, se agradecía al voluntario por la tarea realizada.

Cuando la totalidad de personas habían terminado el etiquetado de *tweets*, para el caso de este estudio participaron doce personas, fueron analizadas las medidas de acuerdo utilizando como medida la Kappa de Fleis.

Segundo etiquetado de *Tweets*

Se presentaron ciertas complicaciones con los modelos de detección de consumo de alcohol en *tweets* y detección de política en *tweets*, los cuales utilizaban como datos de entrada los set de datos generados a partir de los *tweets* etiquetados. Además, como se aprecia en el capítulo siguiente, el resultado de la métrica Kappa de Fleis no fue del todo satisfactoria, en el caso del etiquetado de mención de políticas en *tweets* y en consumo de alcohol en *tweets*. Por esta razón, se tomó la decisión de buscar alguna alternativa factible en cuanto a tiempo y costos para realizar el etiquetado de *tweets*.

Consumo de alcohol

Para algunos temas el etiquetado podría no ser tan fácil de realizar incluso por una persona. El hecho de que tan ambiguo o no es un *tweet* es una condición que presenta ciertas variaciones. En el caso particular del alcohol, para la primera etiqueta, la cual consistía en reconocer si el *tweet* está relacionado con alcohol resulta relativamente fácil. Aunque una de las cosas necesarias era manejar el vocabulario para referirse a las bebidas alcohólicas, las palabras claves estaban disponibles para ser consultada en el sitio web de etiquetado, por lo que esto no generó mayores problemas.

Para el caso de la segunda etiqueta, es decir detectar si el *tweet* hablaba de consumo de alcohol, hubo mucha confusión por todos los matices y variaciones que puede tomar un *tweet* que hable de uso de alcohol. A continuación se muestran algunos ejemplos:

- El *tweet* corresponde a un *Retweet*. En este caso no es totalmente claro que la persona que realizó el *Retweet* comparta la opinión del usuario que generó originalmente el *tweet*, por lo tanto no se le puede atribuir el consumo a la persona que realiza el *retweet*.
- Existe otro grupo de *tweets* que hablan sobre el consumo de alcohol de terceras personas, pero no se puede obtener conclusiones acerca del consumo por parte del autor del *tweet*.
- Existen personas que sugieren el consumo de alcohol a otros usuarios para ciertos fines específicos, por ejemplo en el caso que otro usuario comente que está resfriado o en el caso de bajas temperaturas. Se podría deducir el consumo de alcohol por parte del autor del *tweet* aludiendo a que existe la probabilidad de que haya probado anteriormente estos efectos. Por esta razón es que lo sugiere, pero no es completamente claro.
- Algunos *tweets* no son explícitos. En ciertas oportunidades el usuario manifiesta el deseo de consumir alguna bebida alcohólica, pero no se tiene mayor información si efectivamente llevó a cabo el consumo. A pesar de eso en el caso de sentir la necesidad de consumir una bebida alcohólica, se puede suponer el que autor generalmente hace uso del alcohol.
- En *twitter* se puede observar en reiteradas ocasiones la utilización de la ironía y también del lenguaje implícito en el que es necesario manejar ciertos códigos para entender y estar fami-

liarizado con algunos recursos de *twitter* y las redes sociales como por ejemplo los *hashtags*, ya que ahí se almacena información valiosa del *tweet*.

En el primer etiquetado se dejó que las personas realizaran la tarea libremente, sin antes darle una pauta o fijar las reglas, es decir no se forzó la existencia de acuerdos en el etiquetado. Si bien al principio de la página existían casos que ejemplificaban cada una de las categorías, estos no capturaban todos los matices presentes en los *tweets* de consumo.

Por las razones señaladas anteriormente, no todas las personas estarían capacitadas para realizar el etiquetado. Es necesario estandarizar entre los etiquetadores los criterios que se utilizarán para realizar esta tarea. También es necesario que la persona que realiza el etiquetado tenga una cuenta de *Twitter* y sea un usuario activo de ella, ya sea leyendo frecuentemente los *tweets* que otros usuarios publican o generando contenido.

Para mejorar los rendimientos de los modelos, el etiquetado se realizó de la manera que se explica detalladamente a continuación.

Primero, se hizo una limpieza para eliminar los *Retweets*, para así reducir las ambigüedades asociadas a los retweets que se explicó en los párrafos anteriores.

Se obtuvo una lista aleatoria de *tweets* de los cuales solo fueron guardados los que claramente correspondían a consumo de alcohol o claramente no hablaban de consumo de alcohol por parte del autor del *tweet*. Los *tweets* que presentaran algún grado de ambigüedad fueron descartados. De esta forma se obtuvo el primer etiquetado de *tweets*. Para que este etiquetado fuera válido, la lista de *tweets* ya seleccionados fue ordenada aleatoriamente y se le pidió a una segunda persona de extrema confianza que nuevamente etiquetara todos los *tweets*. Esta persona es un usuario frecuente de las redes sociales, está familiarizado y maneja el lenguaje de *Twitter*, así como sus elementos característicos. Por otro lado conoce y maneja el vocabulario y las expresiones referidas a las bebidas alcohólicas, su consumo, su abuso y sus consecuencias.

De esta forma se obtuvieron 1502 *tweets* etiquetados dos veces. La ventaja de esta forma de etiquetado en comparación al primer etiquetado es que todos los *tweets* fueron revisados por dos personas, y además los criterios para realizar el etiquetado fueron estandarizados. Esto ayuda a que posteriormente los algoritmos no se confundan, ya que de esta forma se evita que el set de datos esté compuesta de información contradictoria. Al tener en un set de datos con dos *tweets* de características similares pero etiquetados en clases distintas, se afecta la precisión de los modelos.

Mención de políticas relacionadas con alcohol.

Para el caso de la tercera etiqueta, es decir identificar si el *tweet* menciona políticas relacionadas con alcohol es necesario que la persona que realizará el etiquetado maneje las políticas y leyes que actualmente rigen en el país en materias de alcohol y que también conozca las nuevas iniciativas que están en la contingencia nacional y despiertan distintas opiniones. Por resumir algunas, se encuentran leyes de impuestos a los alcoholes, penas en el caso que se conduzca bajo la influencia del alcohol y/o que se generen muertes producto de manejar en estado de ebriedad. En este caso, los *tweets* no presentan una amplia ambigüedad como en el caso de los *tweets* de consumo.

Sin embargo, el problema para el análisis de esta etiqueta fue que la proporción de *tweets* re-

lacionados con políticas en la primera muestra de *tweets* era muy baja, del total de 1502 *tweets* solamente 42 *tweets* los etiquetadores consideraron que tenían alguna relación con políticas de alcohol y 1460 *tweets* no hacían alusión a las políticas de alcohol, lo que da como resultado un set desbalanceado.

Un set de datos desbalanceados se refiere a un problema que generalmente se presenta ante un caso de clasificación en donde las clases no están representadas de forma igualitaria, lo que en otras palabras quiere decir que se tiene un número mayor de instancias de una de las clases. El caso particular del etiquetado de políticas relacionadas con alcohol, corresponde a un problema de clasificación de dos clases, osea, un problema de clasificación binario con 1502 instancias o filas. Un total de 42 casos están etiquetados con la Clase Uno, mientas que los 1460 casos restantes fueron etiquetados con la Clase Cero, por lo que se trata de un conjunto de datos desbalanceados y la proporción de la Clase Cero y la Clase Uno es 730:21.

Generalmente, la mayoría de los conjuntos de datos no poseen exactamente el mismo número de instancias de cada una de las clases. En el caso de existir pequeñas diferencias, esto no es de gran importancia, ya que no afecta mayormente el desempeño de los modelos.

En algunos problemas, un desequilibrio en las clases no sólo es algo común si no que corresponde a una situación totalmente esperable. Por ejemplo, en los conjuntos de datos que caracterizan a las transacciones fraudulentas, en donde la mayoría de las transacciones serán de la clase no fraude o los casos de algunas enfermedades en donde la mayoría será etiquetados en el grupo que no posee la enfermedad.

Si bien cada uno de los casos analizados se comporta de manera distinta y no existe una regla estándar que pueda ser tomada como una receta, la literatura advierte que cuando existe un desequilibrio de una proporción 4:1, pueden presentarse problemas. En el caso particular estudiado en este apartado, la proporción 730:21 sobrepasa significativamente la proporción sugerida de 4:1.

Existen diferentes formas de solucionar este tipo de problemas. Uno de ellos es recopilar más datos, en este caso es posible obtener más *tweets* que hablaran de políticas de alcohol considerando que la base de datos de *tweets* está compuesta por 2421554 *tweets* es una alternativa factible poder tener un conjunto de datos con una proporción de clases equilibradas.

Por lo tanto se procedió de manera similar a como se realizó el segundo etiquetado para el caso de los *tweets* que hablaran de consumo de alcohol.

Fue tomada una muestra aleatoria en la que se separaban *tweets* que mencionaran o hicieran alusión a políticas de alcohol y por otro lado se guardaban los *tweets* en los cuales no se nombraban políticas relacionadas con alcohol. De esta forma los 1502 *tweets* fueron revisados y etiquetados por la primera persona. Posteriormente, esta lista fue ordenada aleatoriamente, con el fin de no crear sesgo, y se entregó a una segunda persona que debía leer cada uno de los *tweets* y decidir si a cual de las dos categorías pertenecía. Luego, en el caso que existiera una discrepancia de opinión una tercera persona etiquetó el *tweet* por tercera vez para así poder tener una opinión final solamente en los casos donde la opinión estuviera dividida.

De esta forma se obtuvo un set de datos de 1502 *tweets* etiquetados al menos dos veces.

Es importante destacar que en el caso de *tweets* que se relacionan con políticas no existe un alto grado de ambigüedad como ocurre con el caso de los *tweets* de consumo. Cada uno de los etiquetados plantea desafíos distintos, por lo que las medidas tomadas con el fin de mejorar los modelos deben ser estudiadas caso a caso.

5.3.2. Etiquetado de usuarios

Para obtener esta etiqueta se necesitaba realizar una encuesta directa al usuario. En este caso fue utilizada la encuesta AUDIT, la cual ha sido ampliamente validada y recomendada por el psiquiatra Carlos Ibáñez, y además, es la encuesta que utilizan las distintas organizaciones de salud para medir el consumo de alcohol. La idea es que, a partir de la información entregada por el mismo usuario de *Twitter*, se pudiera clasificar al usuario mediante el puntaje que resultaría de las respuestas. Para esto se necesitó de dos elementos: uno el mecanismo de difusión de le encuesta, por la que se daría a conocer la existencia de esta encuesta a los usuarios de *Twitter* y por otro lado el propio sitio que contendría la encuesta.

Mecanismo de difusión o esparcimiento

Para esto fue implementado un algoritmo haciendo uso de la librería *Twitter4j*. La tarea de este algoritmo era publicar cada cierto tiempo *tweets* mencionando a un usuario en particular, el cual se obtenía de una lista de usuarios chilenos creada con anterioridad y que tenía el nombre de usuario (@usuario). El *tweet* contenía una invitación a contestar la encuesta y la dirección de la url para direccionar al usuario al sitio web de la encuesta. En este algoritmo se incluyó el manejo de credenciales, un elemento crítico cuando se trabaja con *Twitter*, ya que esta red social es extremadamente restrictiva en cuando se trata de cuentas destinadas al spam.

Sitio web de encuesta

En este caso no se utilizó el uso de sesiones como en el caso del sitio web destinado al etiquetado de *tweets*. El sitio utiliza un flujo lineal para mostrar las trece preguntas requeridas. El sitio primero mostraba una página de bienvenida (figura 5.3). A continuación, se mostraba la autorización a un consentimiento informado (figura 5.4) dado la sensibilidad de los datos que el usuario iba a entregar. Es importante este punto para que el usuario tenga la tranquilidad de que la información no será utilizada para un fin distinto al que originalmente fueron recogidos. En el caso de que el usuario aceptara el consentimiento informado, se pasaba a desplegar tres preguntas (figura 5.5). Existían dos casos posibles:

1. El usuario que ingresaba a la encuesta nunca en la vida había consumido alcohol, por lo tanto sólo respondía tres preguntas:
   (a) Edad: existía un campo numérico en donde el usuario ingresaba su edad. El rango de edad permitido era 12 a 90 años.
   (b) Sexo: Hombre o Mujer

(c) cuando había sido la última vez que consumió alguna bebida alcohólica. Las opciones eran las siguientes:
   - Nunca en la vida he consumido alcohol
   - Hace más de un año
   - Hace más de un mes
   - El último mes

2. El usuario que ingresaba a la encuesta había consumido alcohol alguna vez en su vida. En este caso debía contestar las tres preguntas detalladas anteriormente y además, debía responder las diez preguntas correspondientes a la encuesta AUDIT. Estas preguntas eran presentadas en grupos de a dos, con el fin de facilitar el trabajo a la persona encuestada y mejorar la usabilidad (figura 5.6).

Finalmente, en ambos casos se solicitaba al encuestado ingresar su nombre de usuario de *Twitter*, antes de guardar las respuestas y el usuario en la base de datos se validaba que no existieran respuestas previas asociadas a ese usuario. Algunos campos requerían una validación especial, como por ejemplo, que no existieran campos en blanco, es decir que antes de pasar a la siguiente página hubiera respondido todas las preguntas. En el caso de existir errores, éstos se mostraba al usuario para que pudiera continuar con el proceso.

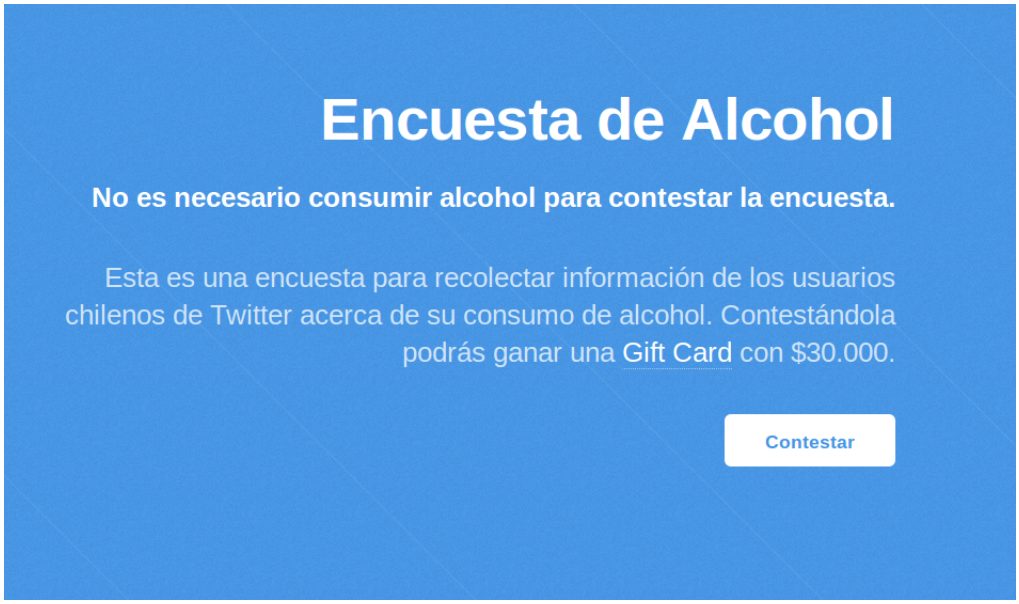

Figura 5.3: Bienvenida del Sitio Web de la encuesta

Fuente: Elaboración propia

5.4. Mantenimiento de datos

Para este trabajo fue necesario utilizar cuatro bases de datos diferentes. La separación de esta bases de datos se debe a que cada una es utilizada en una etapa diferente del proyecto, por lo que se decidió mantenerlas separadas para facilitar su entendimiento.

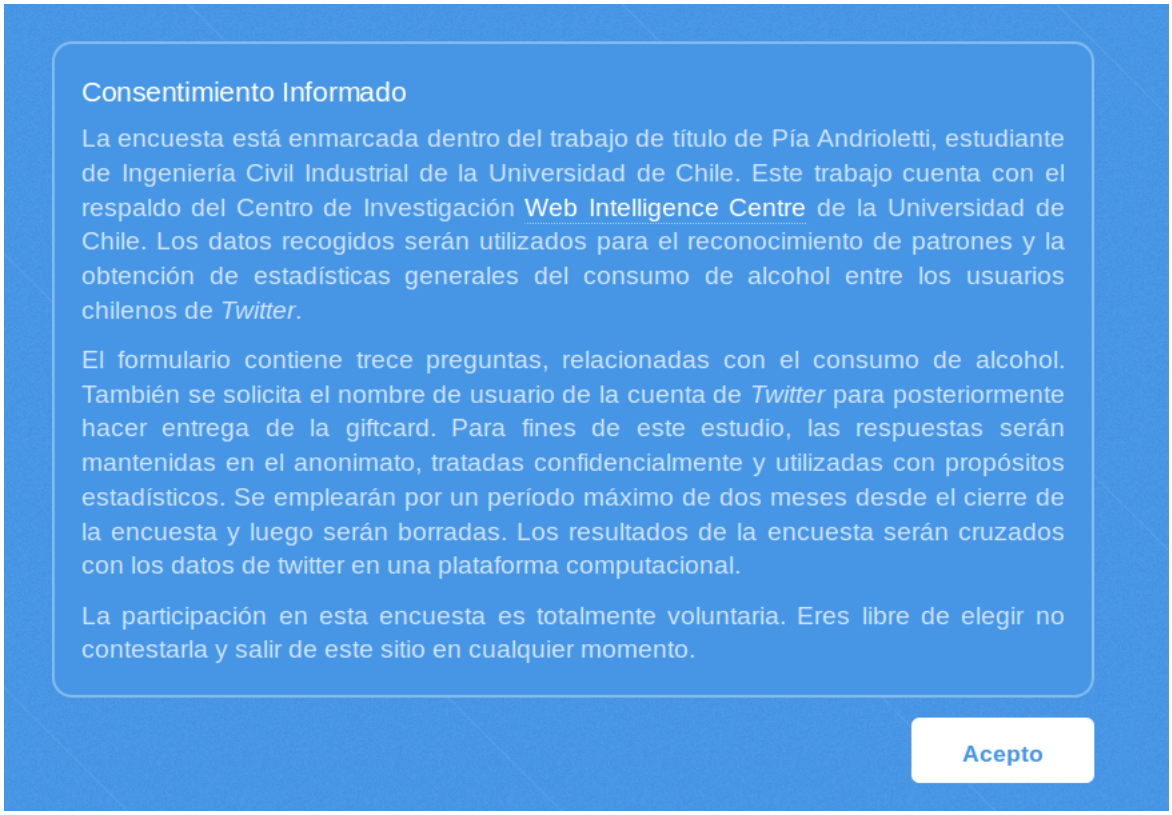

Figura 5.4: Consentimiento informado del Sitio Web de la encuesta

Fuente: Elaboración propia

Tres de las cuatro bases de datos fueron construidas en PostgreSQL y una de ella fue desarrollada en ArangoDB, ya que la estructura y funciones de estas bases de datos facilitaban el trabajo que se realizó con ellas.

En la figura 5.7 se muestra el modelo identidad-relación para la base de datos que fue destinada a guardar los *tweets* y la información de los usuarios chilenos recogidos utilizando la API de *Twitter*.

En la figura 5.8 se muestra el modelo de identidad-relación para la base de datos que fue utilizada con el fin de guardar los etiquetados. En primer lugar, el etiquetado de *tweets* que realizaron los doce voluntarios así como también la base de datos que almacena las respuestas de los usuarios de *Twitter* para hacer el etiquetado de los usuarios.

En la figura 5.9 se observa la base de datos llamada *usertrace* que fue utilizada para crear las métricas de prevalencia y consumo de riesgo y dependencia. *Usertrace* almacena una muestra de usuarios para cada año en base al universo de *tweets* que se tiene para cada año.

Finalmente, en la base de datos construida en ArangoDB se almacenan las relaciones de los usuarios el cual se aprecia en la figura 5.10.

Figura 5.5: Primeras tres preguntas del Sitio Web de la encuesta

Fuente: Elaboración propia

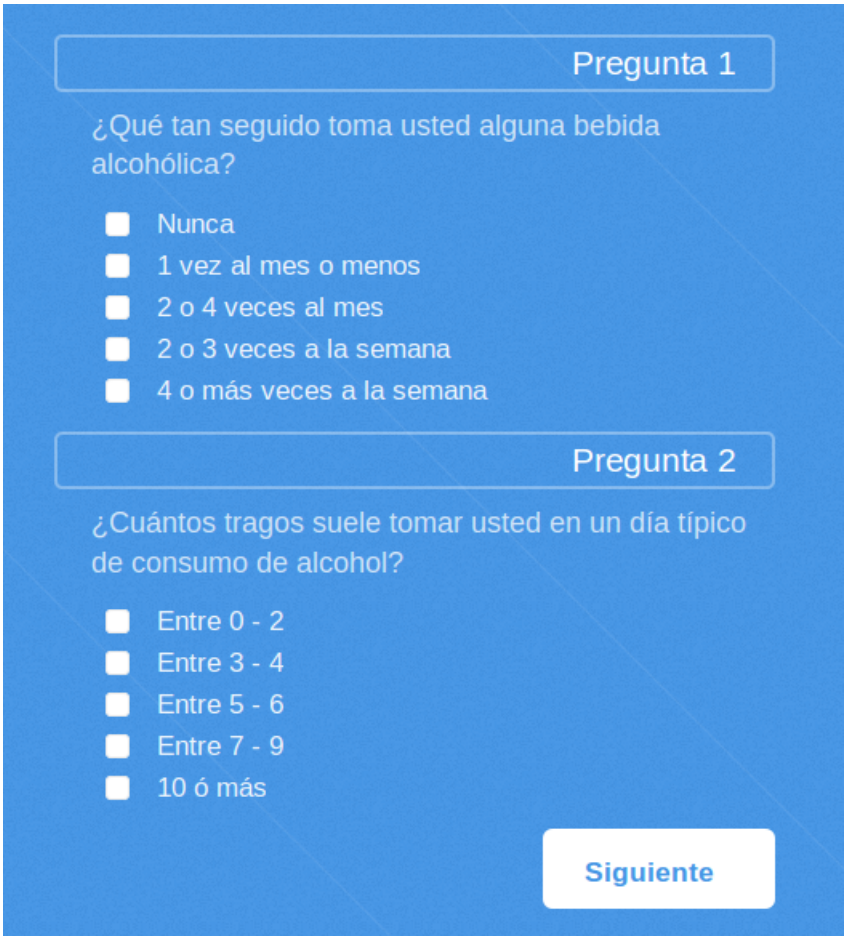

Figura 5.6: Preguntas AUDIT en el Sitio Web de la encuesta

Fuente: Elaboración propia

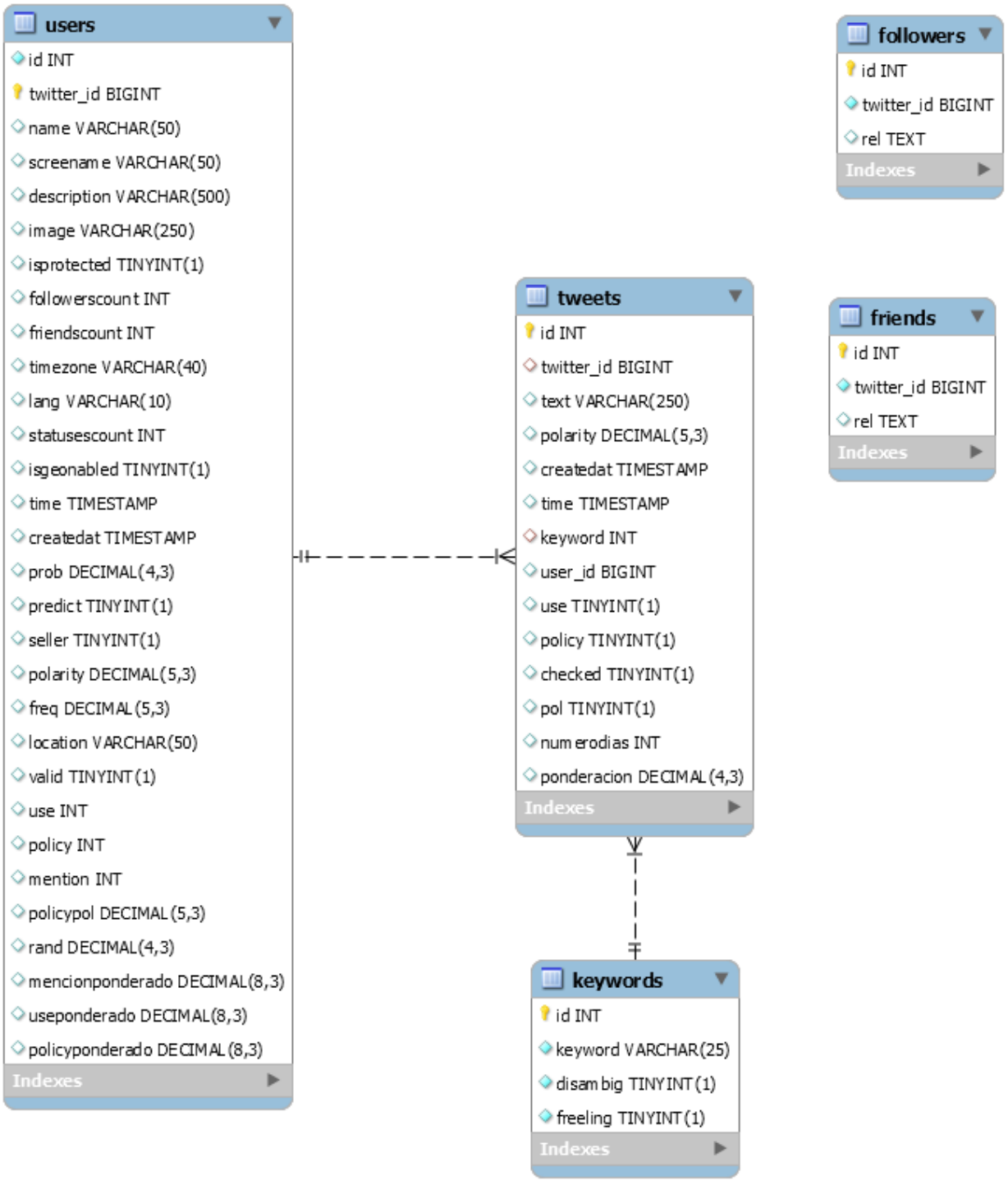

Figura 5.7: Modelo E-R de alcoholdb

Fuente: Elaboración Propia

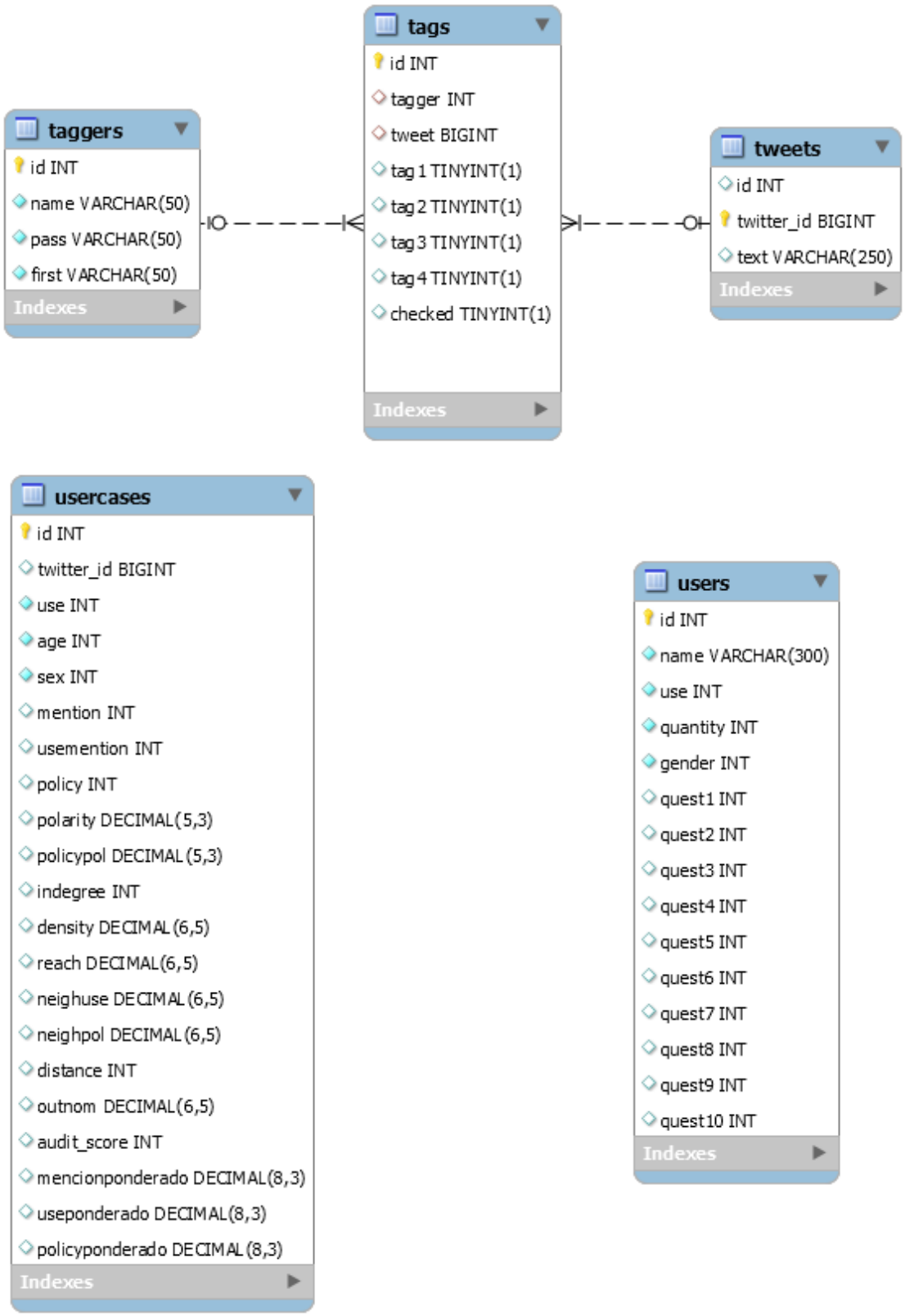

Figura 5.8: Modelo E-R de tagging

Fuente: Elaboración Propia

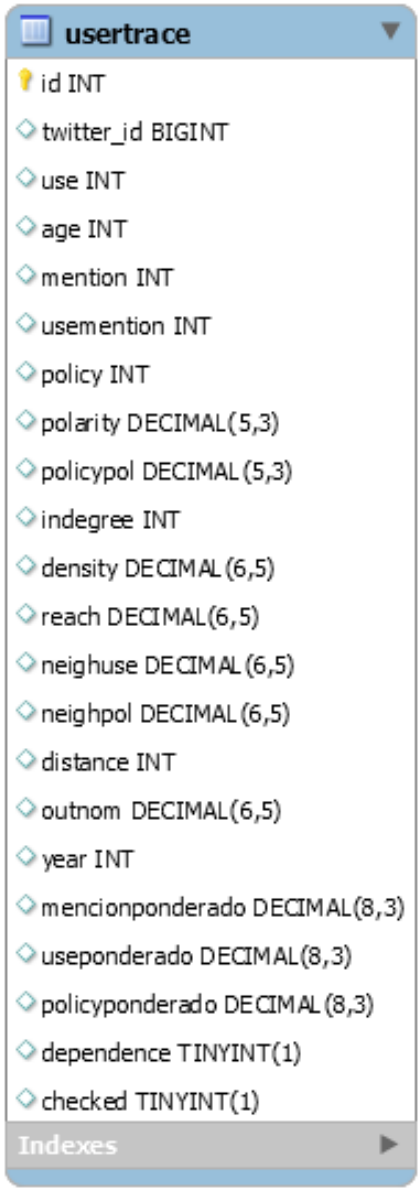

Figura 5.9: Modelo E-R de usertrace

Fuente: Elaboración Propia

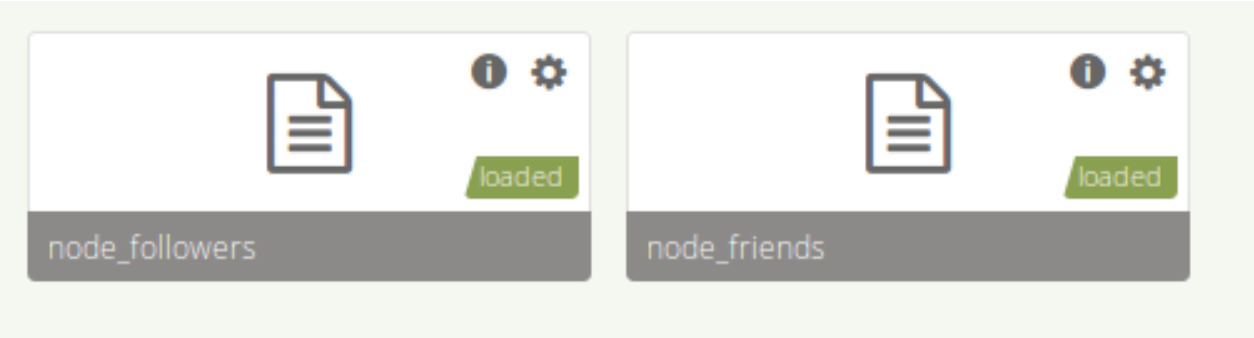

Figura 5.10: Modelo de relaciones utilizando ArangoDB

Fuente: Elaboración Propia

Capítulo 6

Resultados

En este capítulo se presentan los resultados de la implementación descrita en el capítulo anterior. Los resultados presentados a continuación representan información relevante e indispensable para comprender el fenómeno de *Twitter* y el consumo de alcohol englobados en el contexto de las redes sociales.

A continuación, se presentan los resultados de la selección de palabras claves o *Keywords*, las que son necesarias para saber que *tweets* recolectar en el universo de *Twitter*.

6.1. Palabras claves o Keywords

De la encuesta realizada se obtuvieron en total 283 respuestas, lo que finalmente se tradujo en 717 términos que hacen referencia a: el consumo de alcohol, a sus efectos, a las personas que lo consumen o a los estados en que se puede encontrar una persona que consume alcohol, a distintas marcas asociadas a bebidas alcohólicas o a distintas combinaciones para preparar bebidas alcohólicas.

Algunas palabras que presentaban ambigüedad se utilizó el desambiguador "tom", el que hace referencia al verbo tomar, se utilizó esta palabra ya que es la forma como es consumido generalmente el alcohol y también es de las palabras más usadas para señalar que se va a beber alcohol. Un ejemplo de estas palabras ambiguas es terremoto, la que hace alusión a los movimientos telúricos muy comunes en nuestro país y también a una bebida alcohólica preparada con helado de piña y pipeño.

El proceso finalmente concluyó en un total de 103 cadenas de caracteres. Algunos de ellos son simples, otros son bigramas, trigramas y cuatrigramas.

Es importante destacar que estas *keywords* son válidas solamente para un contexto acotado a Chile, ya que un porcentaje importante de ellas son modismos propios de nuestro país o son palabras asociadas a marcas que podrían no ser comercializadas en otros países.

En la tabla 6.1 se encuentra la lista final de palabras claves, y también se especifica entre paréntesis el desambiguador, en los casos en que se tuvieron que utilizar.

6.2. Recolección de datos

Se obtuvieron en total 14738317 cuentas de usuarios, de las cuales 14588785 son válidas, mientras que 149532 no lo son.

Del total de cuentas obtenidas, se observa que 149340, es decir un 9,11 %, están protegidas, por lo tanto no fue posible acceder a la información generada por estos usuarios. Por otro lado, 1488977 (un 90,88 %) son cuentas públicas, que el porcentaje de cuentas públicas sea alto es de vital importancia en este estudio, porque ayuda a que se pueda acceder a la información. Esta relación se puede apreciar con mayor claridad en el gráfico de la figura 6.1.

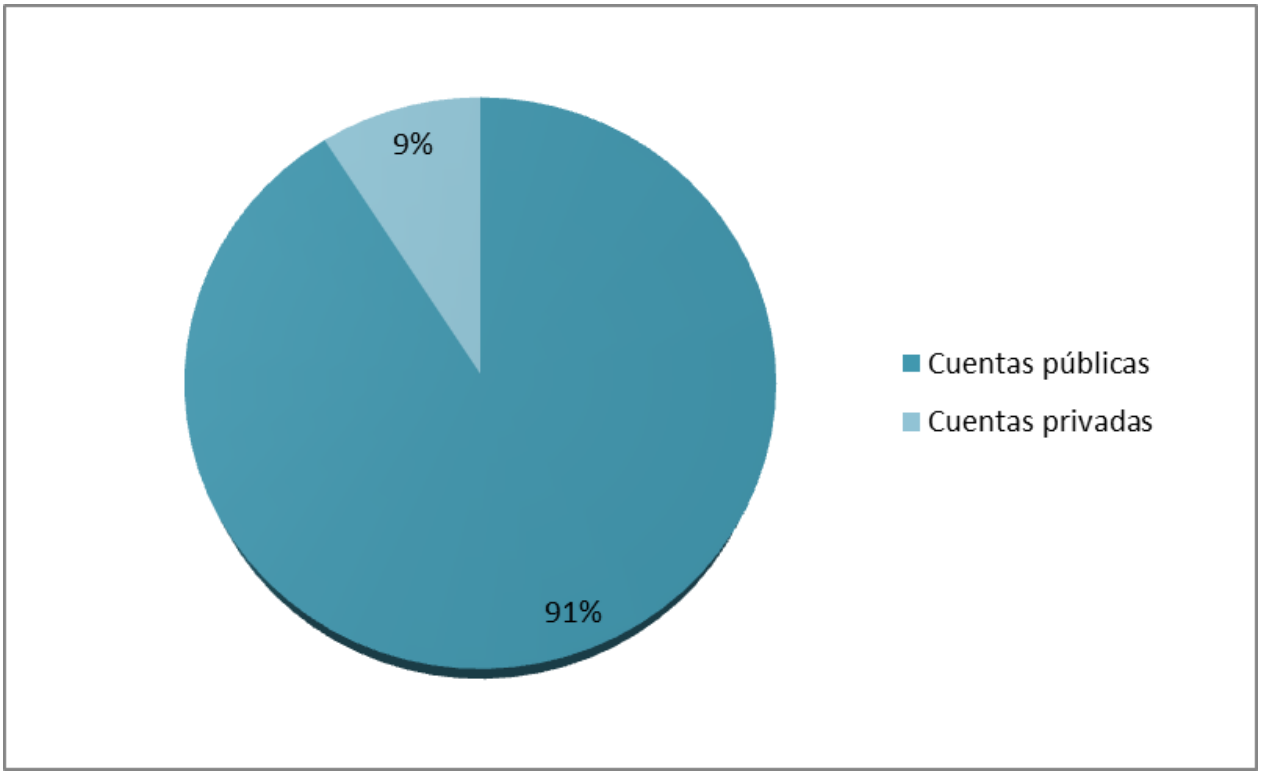

Figura 6.1: Porcentaje de cuentas públicas y privadas en *Twitter*

Fuente: Elaboración propia

En la tabla 6.2 se observa el comportamiento que ha tenido el crecimiento en el número de cuentas de usuario chilenos en *Twitter*. En el gráfico de la figura 6.2 se puede ver el número de cuentas acumulado para cada año entre el 2006 y el 2017.

Términos Utilizados		
aguardiente	alcohol	baltica
baltiloca	birra	capel
cervecita	cerveza	chela
chelits	combinao	copete
copetito	copetits	cufifo
doragua	ebrio	fernet
guarisnaque	licor	mojito
navegao	pilsen	pisco
roncola	tequila	tinto
tomatera	traguito	unas heladas
vinacho	vinito	vituperio
vodka	whiscola	whiskacho
wiscacho	whisky	trago
espumante	empinar el codo	rayuela corta
xelita	ron	chelita
chocoron	champagne	copetiwi
baltiloka	ponche	michelada
copetits	Baileys	curao
arriba de la pelota	roncito	pisquito
piskito	wiscola	balticrazy
stella artois	pipeño	tintito
tequiliwis	borrachera	champaña
schop	chupilca	vino (freeling)
copita (tom)	cortito (tom)	malicia (tom)
puritano (tom)	terremoto (tom)	jote (tom)
cusqueña (tom)	budweiser (tom)	becker (tom)
corona (tom)	heineken (tom)	dorada (tom)
escudo (tom)	royal guard (tom)	mistral (tom)
chicha (tom)	malbec (tom)	petit verdot (tom)
sauvignon blanc (tom)	chardonnay (tom)	riesling (tom)
cabernet sauvignon (tom)	merlot (tom)	carmenere (tom)
syrah (tom)	pinot noir (tom)	bauza (tom)
havana club (tom)	jack daniel (tom)	mitjans (tom)
Johnnie Walker (tom)	cola de mono (tom)	colemono (tom)
un pencazo (tom)		

Tabla 6.1: Tabla de términos utilizados.

La etiqueta (tom) indica la aplicación del desambiguador. La etiqueta (freeling) indíca que en esa palabra clave fue utilizada la regla de desambiguación de PoS Tagger.

Fuente: Elaboración Propia

Año	Número de cuentas nuevas	Número de cuentas acumulado
2006	39	39
2007	1813	1852
2008	6240	8092
2009	150538	158630
2010	394802	553432
2011	317533	870965
2012	210027	1080992
2013	132335	1213327
2014	98369	1311696
2015	109323	1421019
2016	216949	1637968
2017	349	1638317

Tabla 6.2: Número de cuentas nuevas de usuarios chilenos creadas por año y número acumulado de cuentas de usuarios chilenos.

Fuente: Elaboración Propia

Año	Número de *tweets*
2007	97
2008	489
2009	7828
2010	76393
2011	240982
2012	430439
2013	444873
2014	356536
2015	374129
2016	477293
2017	12495

Tabla 6.3: Número de *tweets* por año que contienen las keywords seleccionadas.

Fuente: Elaboración Propia

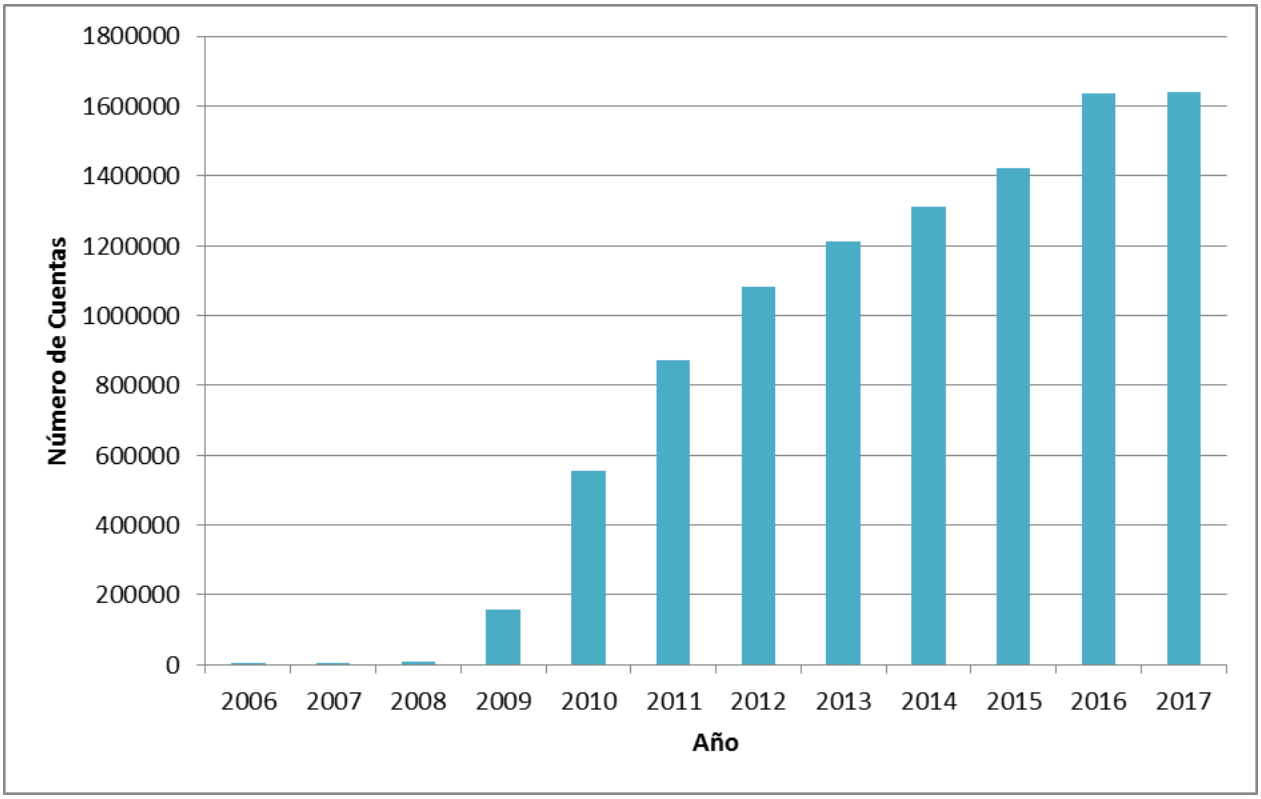

Figura 6.2: Número de cuentas creadas total acumulado por año

Fuente: Elaboración propia

6.3. Etiquetado

En esta sección se muestran los resultados de los dos procesos de etiquetados que fueron necesarios realizar en este trabajo para posteriormente poder hacer el entrenamiento de los algoritmos. Primero, se presentan los resultados del etiquetado de *tweets*. Posteriormente, se presentan los resultados de la encuesta realizada a los usuarios. Los resultados obtenidos de estos dos procesos de etiquetado condicionan los resultados y el desempeño que tengan los modelos construidos.

6.3.1. Etiquetado de *tweets*

Del proceso de recolección de *tweets* se obtuvieron 2421554 *tweets* que contienen las palabras claves seleccionadas. A partir de este número, se calculó el tamaño de la muestra adecuado, tomando en consideración que por una parte se desea tener representada en esta muestra el universo de *tweets* extraídos, pero por otro lado no es factible el etiquetado de una gran cantidad de *tweets* ya que el etiquetado es una tarea tediosa en términos de recursos y tiempo. Por lo tanto para obtener el tamaño de la muestra se consideró:

- Tamaño del universo: 2421554
- Heterogeneidad: 50 %

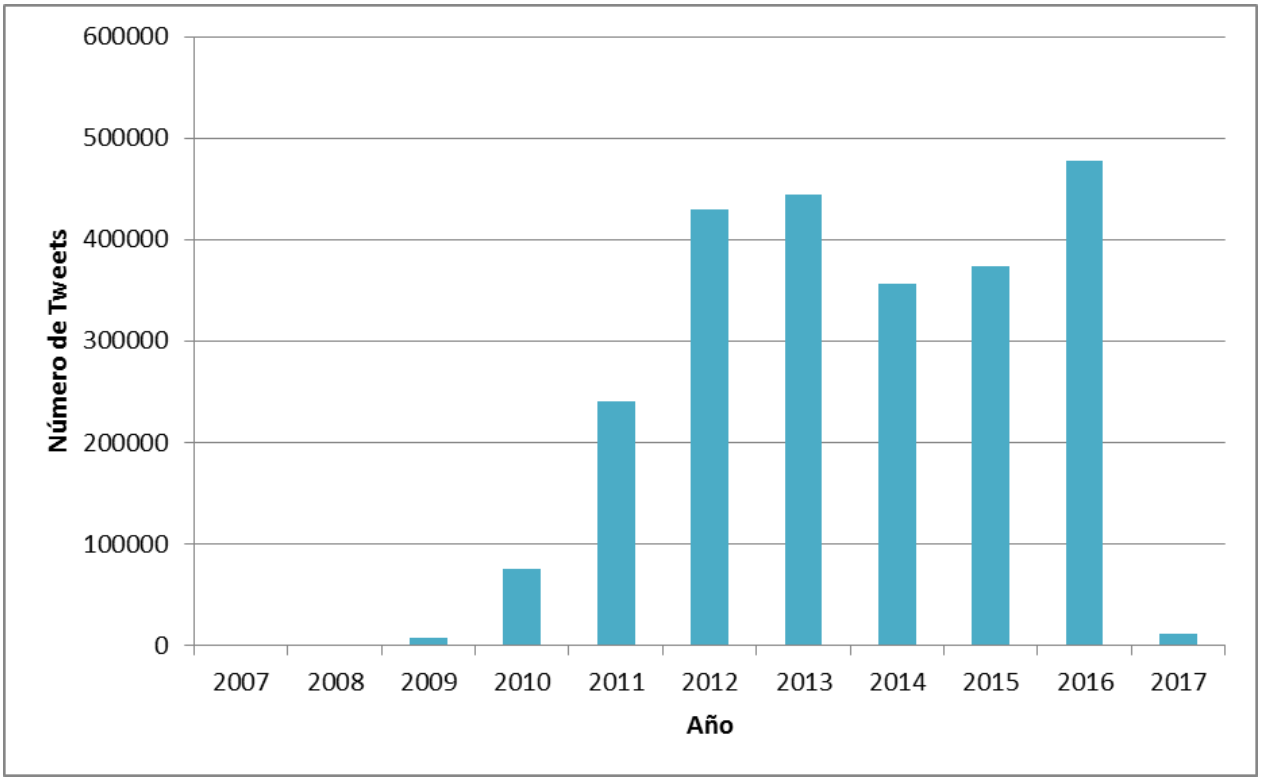

Figura 6.3: Número de *tweets* relacionados con alcohol por año.

Fuente: Elaboración propia

- Margen de error: 3 %
- Nivel de confianza: 98 %

El tamaño de la muestra resultante fue de 1502 *tweets*, que fueron extraídos aleatoriamente de la base de datos con los *tweets* que contenían solamente *tweets* con las palabras claves que fueron nombradas anteriormente.

Se tomó una muestra aleatoria de 1502 *tweets*, considerando las fórmulas para obtener una muestra representativa de todos los *tweets* extraídos desde la red social.

A continuación se muestran los resultados obtenidos del primer etiquetado de *tweets*. Como se mencionó anteriormente, se solicitó a los voluntarios etiquetar los *tweets* en cuatro categorías distintas: la primera permitiría conocer cuál es la precisión de las palabras claves elegidas para la recolección de *tweets* desde *Twitter* y las otras tres categorías serían usadas para el entrenamiento de algoritmos.

En el proceso de etiquetado participaron doce personas. De la muestra de 1502 *tweets*, cincuenta *tweets* fueron etiquetados por todas las personas que participaron en el proceso (osea, cada uno de estos cincuenta tweets fueron etiquetados doce veces). Por otro lado, cada una de las doce personas etiquetó además ciento veintiún *tweets*, dando como resultado 1452 *tweets* etiquetados solamente una vez.

En este punto se analizará las medidas de acuerdo entre las doce personas, puesto que la idea es

cuantificar las diferencias de opinión entre los cincuenta *tweets* etiquetados por todos ellos.

Categoría	**Homogeneidad**
Ligado a alcohol	91,15 %
Consumo de alcohol	49,60 %
Políticas relacionadas con alcohol	2,80 %
Venta de alcohol	3,40 %

Tabla 6.4: Heterogeneidad en las etiquetas

Fuente: Elaboración Propia

Categoría	**Acuerdo Relativo**	**Kappa de Fleiss**
Ligado a alcohol	0,97	0,76
Consumo de alcohol	0,78	0,56
Políticas relacionadas con alcohol	0,98	0,39
Venta de alcohol	0,98	0,79

Tabla 6.5: Medidas de acuerdo en el primer etiquetado.

Fuente: Elaboración Propia

6.3.2. Etiquetado de usuarios

Como ya se ha dicho anteriormente, para poder construir el clasificador de usuarios fue necesario etiquetarlos. Para este fin, se difundió una encuesta que se llevó a cabo desde el día miércoles 23 de noviembre del 2016 hasta el día lunes 6 de marzo del 2017.

La encuesta fue enviada a 6800 cuentas de *Twitter*, y se obtuvieron 149 respuestas, lo que se traduce en una tasa de respuesta del 2,19 %.

Posteriormente, al momento de cruzar los datos con la base de datos utilizada en este estudio, la cual contenía a los usuarios chilenos, quedó un total de 121 respuestas válidas con las cuales se trabajó y fueron construidos los algoritmos.

El resto de las respuestas correspondía a usuarios que no fueron encontrados en la base de datos, por distintos motivos. Algunos no están integrados a la red de usuarios chilenos construida para este trabajo y otros corresponden a nombres de usuarios que ni siquiera se encuentran registradas en *Twitter*, lo cual se comprobó haciendo una simple búsqueda de forma manual utilizando la herramienta de buscador de *Twitter*. Esto último se debe a que cuando se distribuye una encuesta por internet existe el problema del anonimato. A pesar de que al final de la encuesta se le pedía a la persona registrar su nombre de usuario, la página diseñada no contaba con un mecanismo de verificación de la identidad del usuario que contestaba la encuesta, por lo que en el campo del nombre de usuario, alguien podía ingresar cualquier cadena de caracteres.

Para asegurarse de que todo estuviera en orden, se realizó una revisión de las respuestas que

había resultado del cruce con la base de datos, con el fin de eliminar posibles valores atípicos[1].

De esta revisión se encontraron dos casos inválidos. El primero era el caso de un nombre de usuario con el mismo nombre de un canal de televisión, que era el ejemplo para llenar el campo de usuario. Como esta cuenta representa a una institución y no a una persona, este dato fue descartado, ya que los análisis que derivan de este usuario no pueden ser considerados útiles ni válidos. El segundo caso correspondía a alguien que tenía el puntaje máximo en la encuesta AUDIT (cuarenta puntos), pero que solo tenía doce años de edad. Este dato fue revisado, y como en la descripción de la cuenta de este usuario, él se describía como una persona de 25 años, no se consideró este registro, ya que una de las variables que consideran los modelos es la edad de la persona, por lo que podría traducirse en perturbaciones a los modelos.

Con las trece preguntas de la encuesta fueron analizadas ciertas estadísticas relacionadas al consumo de alcohol. Una de ellas es la distribución de la edad de los encuestados. En el gráfico 6.4 se observa que no fueron recopilados datos de usuarios cuyas edades estén en los grupos extremos. El intervalo con mayor presencia es entre los 21 y los 30 años, la mediana de la muestra corresponde a 27 años y el promedio se encuentra en 29,35 años. Este punto va de acuerdo con la creencia de que los usuarios de *Twitter* son en promedio más jóvenes que la población general que habita el país.

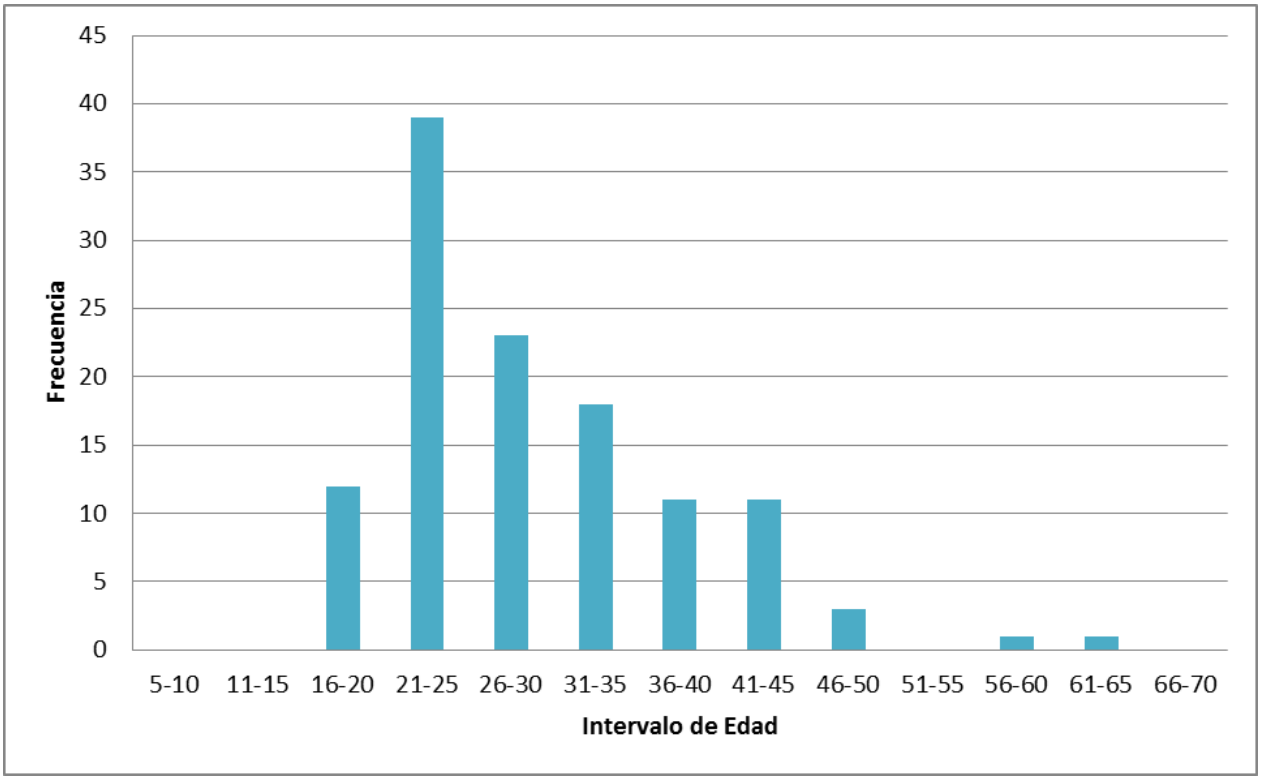

Figura 6.4: Distribución de edad de la muestra

Fuente: Elaboración propia

Como se puede apreciar en la tabla 6.6 y en el gráfico 6.5, el 3 % de los encuestados respondió no haber consumido nunca alcohol. Por otra parte el 97 % reconoció haber consumido alguna bebida

[1]En inglés *outlier*

alcohólica en su vida, el 92 % dice haber consumido alcohol al menos una vez en el último año y el 82 % consumió alcohol el último mes.

Cabe hacer notar el bajo porcentaje de personas que nunca han consumido una bebida alcohólica. Es probable que los altos porcentajes de prevalencia reflejados en el gráfico 6.5 se deba a que la encuesta presente cierto sesgo. A pesar que se hizo hincapié en que la encuesta no estaba destinada exclusivamente a personas que consumen alcohol, existe la posibilidad que haya sido considerada de esa forma. Esto se vio reflejado en que algunos usuarios manifestaron su malestar al recibir la encuesta pensando que ésta era enviada a personas con consumo excesivo de alcohol.

Prevalencia	**Porcentaje**
Prevalencia de vida	97 %
Prevalencia anual	92 %
Prevalencia del mes	82 %

Tabla 6.6: Porcentaje de prevalencia de la muestra.

Fuente: Elaboración Propia

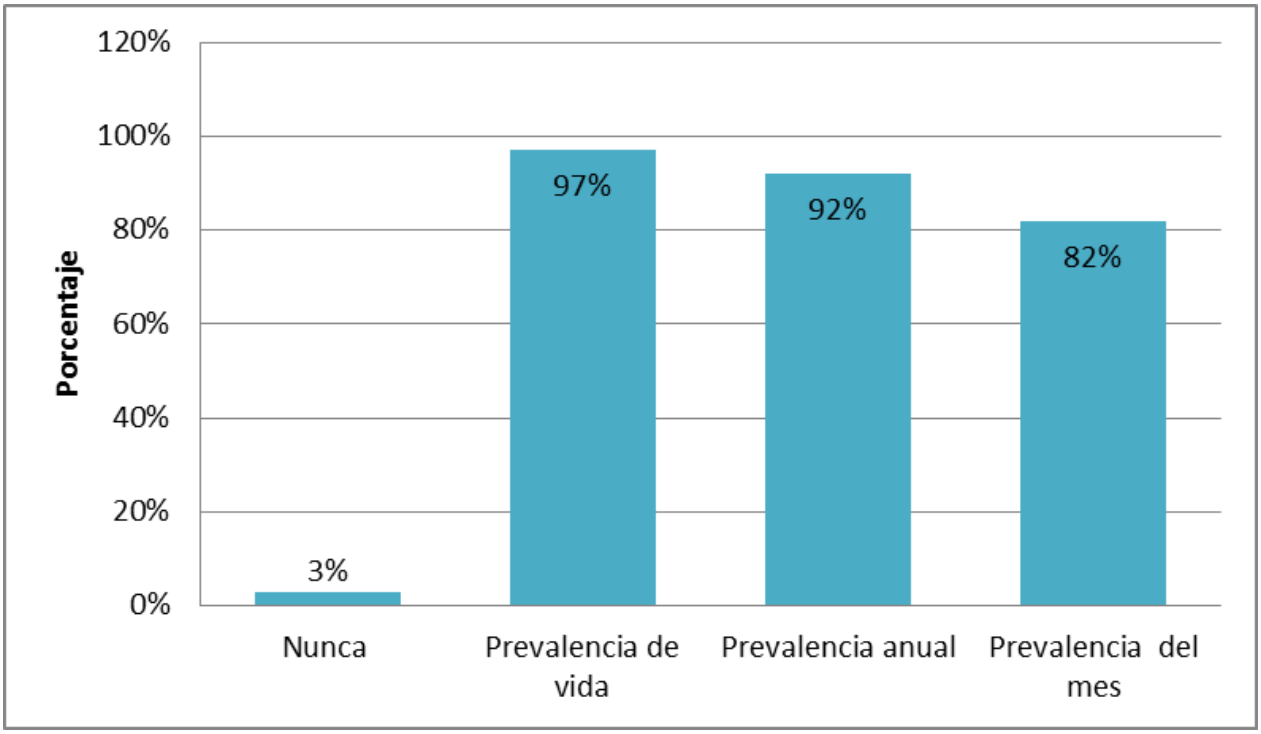

Figura 6.5: Distribución de la pevalencia de la muestra

Fuente: Elaboración propia

Por otra como se puede apreciar en el gráfico de la figura 6.6 el 70 % de las encuestas recibidas fue respondida por usuarios de sexo masculino.

De los datos obtenidos de la encuesta se sugiere que existe en *Twitter* un predominio de usuarios de sexo masculino, lo cual se diferencia de las cifras que entrega el CENSO de nuestro país del año 2012, en donde la población femenina corresponde al 51,4 %. Esto va de acuerdo con la creencia

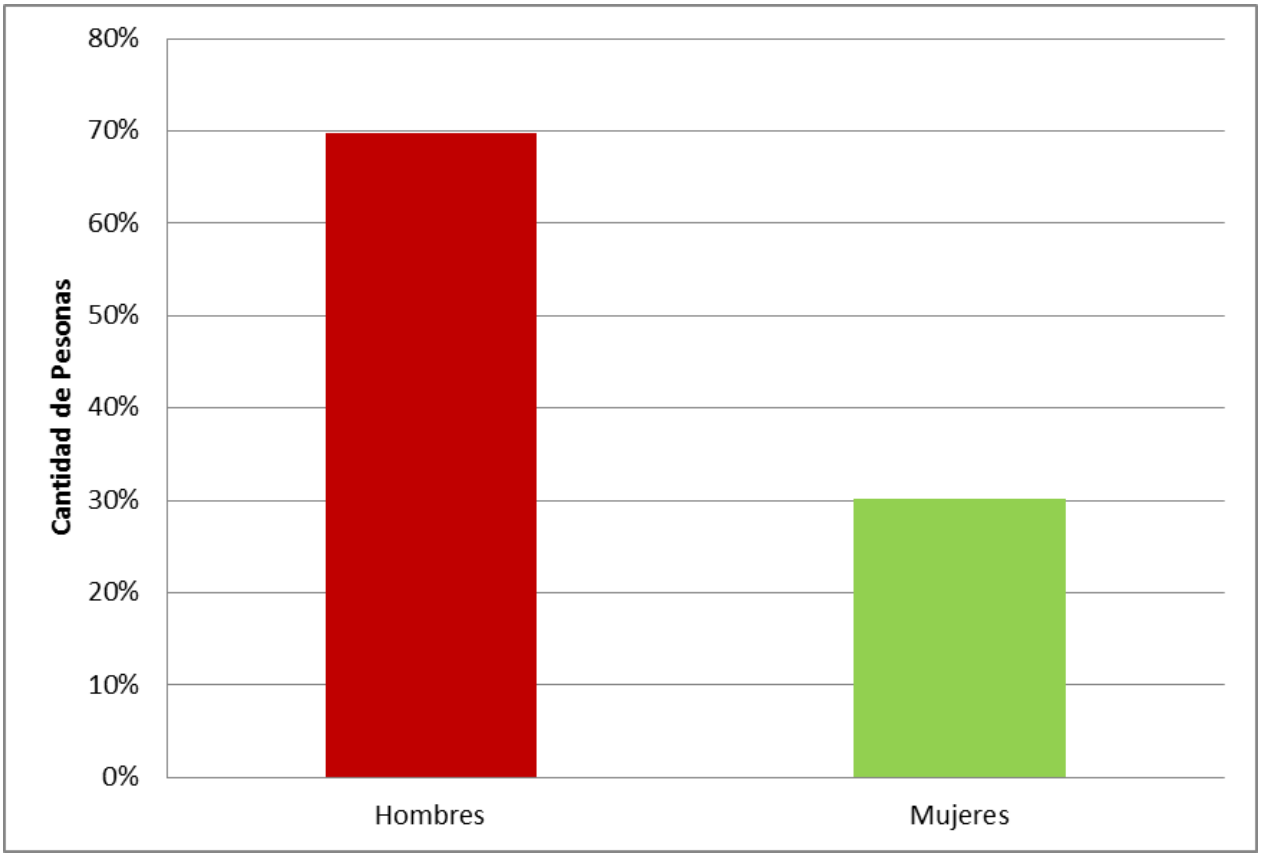

Figura 6.6: Distribución de lo encuestados según sexo

Fuente: Elaboración propia

de que en *Twitter* existe una mayor presencia masculina. La razón de la diferencia que existe entre los datos recogidos de la encuesta y la del CENSO podría encontrarse en que las mujeres prefieren ser parte de otras redes sociales distintas de *Twitter*, ya que podrían existir elementos que hacen que *Twitter* sea más popular entre hombres.

Es importante destacar que existe la posibilidad de que la encuesta esté sesgada. Para evitar esto, la encuesta fue distribuida de forma aleatoria a partir de una lista que contenía los nombres de usuarios chilenos de *Twitter* obtenida de la base de datos generada. Una explicación de esto puede ser que los usuarios al recibir el *tweet* con la invitación a participar de la encuesta pensaron que estaba destinada exclusivamente a personas que consumen habitualmente alcohol, a pesar de que al momento de distribuir la encuesta se destacó que podía ser contestada por cualquier persona, independiente de si consume alcohol o no.

Como se habló en los capítulos anteriores, la prevalencia de consumo de alcohol predomina mayormente en la población masculina. Según el Décimo Primer Estudio de Drogas en Población General en el año 2014 la prevalencia de consumo de alcohol en el último mes en el caso de los hombres fue de 55,3 % y en el caso de las mujeres fue de 42,5 %. Históricamente, el alcohol ha tenido un predominio en el sexo masculino como se ve en el gráfico 6.8. Este hecho pudo haber incidido en obtener un mayor número de respuestas de parte de hombres. Otra explicación se podría encontrar en que las mujeres que utilizan la red social *Twitter* tienen más recelo a entregar

información de carácter personal y sensible o son más desconfiadas en las plataformas de las redes sociales.

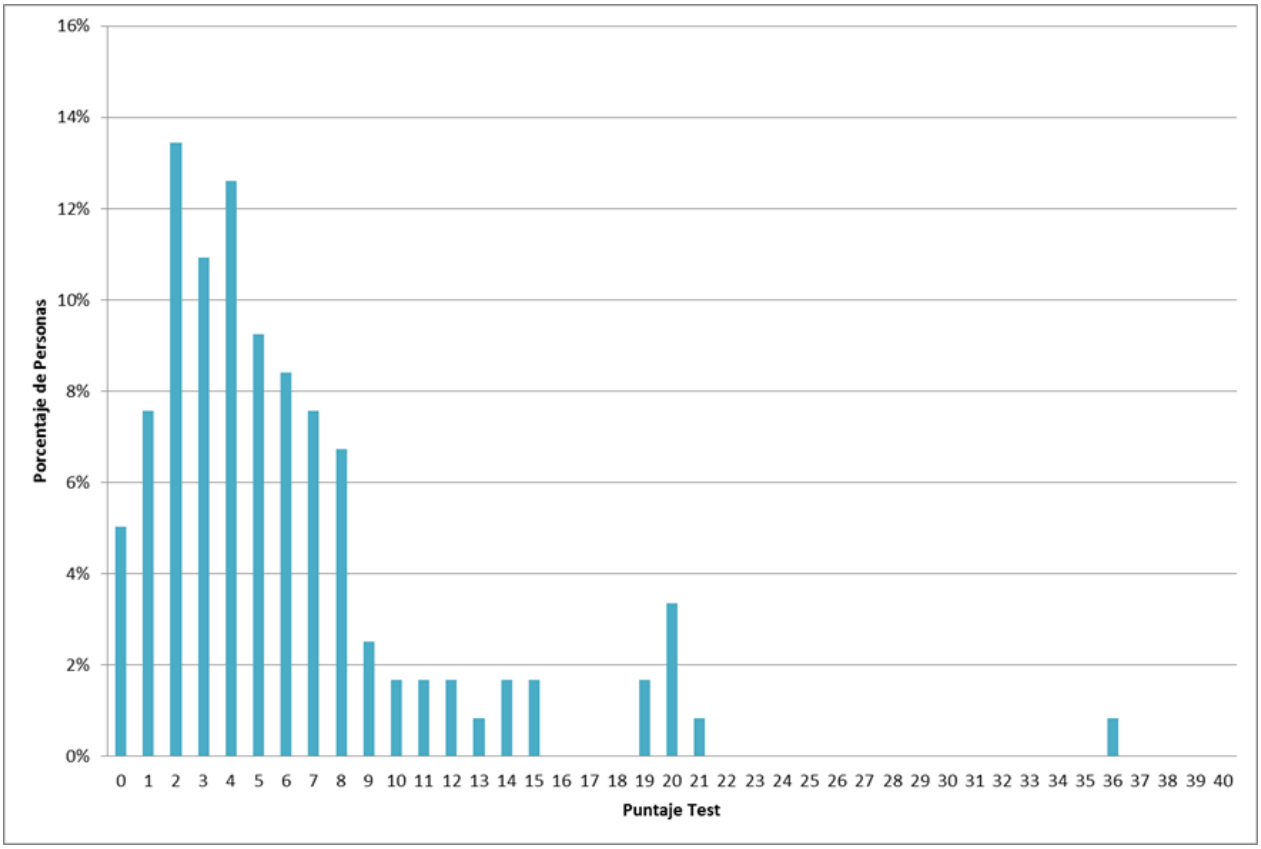

Figura 6.7: Distribución de la muestra según puntaje AUDIT

Fuente: Elaboración propia

Por otra parte el 73,554 % de los encuestados obtuvo un puntaje menor a 8 puntos en la encuesta AUDIT, mientras que un 26,446 % obtuvo un puntaje igual o mayor a ocho puntos en la encuesta AUDIT.

La distribución con respecto a los puntajes de la escala AUDIT muestra que el 56,64 % de la población encuestada se encuentra bajo los 6 puntos, es decir, bajo el umbral de riesgo considerado por la definiciones de la Organización Mundial de la Salud y por la validación para Chile.

De acuerdo a las categorías de riesgo propuesta por la Organización Mundial de la Salud, los resultados obtenidos se muestran en la tabla 6.7. Por lo tanto, se tiene que el 26,55 % de quienes han consumido alcohol en el último año presentan una puntuación AUDIT superior al límite de 8 o más puntos, puntaje de corte definido por la OMS.

Por otro lado, al utilizar la escala definida por la validación realizada en Chile, se obtienen los resultados mostrados en la tabla 6.8. De esta tabla, se puede traducir que el 43,37 % de la población considerada se encuentra por sobre el umbral de riesgo, es decir seis o más puntos, según la definición dada por la validación chilena del instrumento.

Figura 6.8: Prevalencia de alcohol por mes separado según sexo

Fuente: Senda

6.4. Evaluación de algoritmos

En esta sección se muestra el desempeño de los modelos construidos. Posteriormente, en base a esto se elige el algoritmo que posea el mejor desempeño.

Clasificación OMS	**Puntuación**	**Porcentaje del total sobre prevalentes del último año**
Consumo de Riesgo.	Entre 8 y 15 puntos	19,47 %
Consumo Perjudicial.	Entre 16 y 19 puntos	1,77 %
Dependencia de Alcohol.	20 puntos o más	5,31 %

Tabla 6.7: Escala AUDIT (OMS).

Fuente: Elaboración Propia.

Validación Chile	Puntuación	Porcentaje del total sobre prevalentes del último año
Consumo de Riesgo	Entre 6 y 8 puntos	23,90 %
Consumo Perjudicial/Dependencia	9 o más puntos	19,47 %

Tabla 6.8: Escala AUDIT Validación Chile.

Fuente: Elaboración Propia.

6.4.1. Detección de consumo en *tweets*

Para poder categorizar si un *tweet* está relacionado con consumo de alcohol por parte de quien emite el *tweet* se utilizó un modelo de categorización binaria, es decir un problema de clasificación simple. En este caso se trabajó con cuatro algoritmos: *Naive Bayes*, *Voted Perceptron*, árboles de decisión y *Support Vector Machines*.

Se obtuvieron varios modelos, variando los parámetros de estos algoritmos, y a continuación se presentan los que presentaron un mejor desempeño. Por otra parte, como se mencionó anteriormente, se llevaron a cabo dos instancias de etiquetado con metodologías distintas, por lo que se hace una comparación para analizar cuanto influye el etiquetado.

En la tabla 6.9, se muestran los resultados del mejor modelo para el primer etiquetado. Por otro lado, en la tabla 6.10 se muestra el rendimiento del mejor modelo para el segundo etiquetado.

Clase	*Precision*	*Recall*	F-Measure	ROC Area
No consumo (0)	0,677	0,587	0,628	0,678
Consumo (1)	0,621	0,707	0,662	0,678
Ponderado	0,650	0,646	0,645	0,678

Tabla 6.9: Rendimiento de *Naive Bayes* para la detección de consumo en el primer etiquetado

Fuente: Elaboración Propia

Clase	*Precision*	*Recall*	F-Measure	ROC Area
No consumo (0)	0,874	0,790	0,830	0,849
Consumo (1)	0,842	0,907	0,873	0,849
Ponderado	0,856	0,855	0,854	0,849

Tabla 6.10: Rendimiento de *SVM* para la detección de consumo en el segundo etiquetado

Fuente: Elaboración Propia

Se observa que, gracias al segundo etiquetado se obtuvo una importante mejora en toda las métricas del segundo modelo. Esto es transcendental para asegurar la confiabilidad del modelo de consumo de alcohol en usuarios.

6.4.2. Detección de políticas en *tweets*

Para construir este modelo, al igual que en el caso anterior, la mención de políticas relacionadas con alcohol en los *tweets* fue tratado como un problema de clásico de clasificación binaria.

En esta oportunidad fueron utilizados los mismos cuatro algoritmos mencionados en la detección de consumo: *Naive Bayes*, *Voted Perceptron*, árboles de decisión y *Support Vector Machines*.

A continuación se presentan los modelos con mejor desempeño para los dos etiquetados.

En la tabla 6.11 se muestran los resultados del mejor modelo para el primer etiquetado. Por otro lado, en la tabla 6.12 se muestra el rendimiento del mejor modelo para el segundo etiquetado.

Clase	***Precision***	***Recall***	**F-Measure**	**ROC Area**
No políticas (0)	0,979	0,942	0,960	0,776
Políticas (1)	0,124	0,286	0,173	0,778
Ponderado	0,955	0,923	0,938	0,776

Tabla 6.11: Rendimiento de *SVM* para la detección de políticas en el primer etiquetado

Fuente: Elaboración Propia

Clase	***Precision***	***Recall***	**F-Measure**	**ROC Area**
No políticas (0)	0,920	0,985	0,951	0,933
Políticas (1)	0,977	0,880	0,926	0,933
Ponderado	0,944	0,941	0,941	0,933

Tabla 6.12: Rendimiento de *SVM* para la detección de políticas en el segundo etiquetado

Fuente: Elaboración Propia

En este caso, el problema que se presentó en la construcción del primer modelo, en donde se utilizaron los datos resultantes del primer etiquetado, fue que la clase de interés, es decir, la clase 1 tenía un bajo porcentaje de datos. Por esa razón, en la tabla 6.11 se aprecia que las métricas para la clase de interés son demasiado bajas.

Al crear el segundo modelo, mostrado en la tabla 6.12, este resultó en una mejora considerable para la clase de interés. Si bien la *Precision* ponderada muestra una baja, debido a la baja en la *Precision* de la clase cero, este aspecto negativo es compensado por un aumento en la *Precision* y *Recall* de la clase uno, además de un aumento en el *Recall* de la clase cero.

Realizar el segundo etiquetado para los *tweets* de políticas fue de gran ayuda para mejorar el desempeño del modelo.

6.4.3. Consumo de alcohol en usuarios

La clasificación de consumo de alcohol presente en los usuarios de *Twitter* fue modelada como un problema de Minería de Datos (en esta ocasión no será realizado un tratamiento de textos para conseguir el conjunto de atributos). Para esto fueron utilizados dieciséis atributos, los cuales están compuestos por medidas calculadas a partir de los *tweets*, el entorno social del usuario y características propias de la persona que posee la cuenta.

Para la construcción del modelo de consumo de alcohol en usuarios, se utilizaron tres algoritmos con el fin de poder comparar los rendimientos de éstos y así elegir el mejor. Los algoritmos utilizados fueron: *Support Vector Machine*[2], *Multilayer Perceptron*[3] y *Voted Perceptron*. Para los tres algoritmos, las medidas de rendimiento obtenidas fueron muy similares, variando levemente en la *Precision* para la clase 0, es decir la clase que no consume alcohol. En otras palabras, los tres algoritmos extraen todo el poder predictivo disponible que permite el conjunto de variables elegidas.

Dada la naturaleza de los datos obtenidos, como ya se vio en la figura 6.5 se tomó como clase consumidora a aquellas personas que habían consumido alcohol el último mes, esto con el fin de tener una proporción adecuada para que el algoritmo pudiera aprender de los datos. En esta oportunidad los datos presentaron un predominio de la clase de interés, es decir la clase consumidora de alcohol.

Nuevamente el algoritmo que obtuvo un mejor rendimiento es el Support Vector Machine (SVM). En la tabla 6.13 se muestran las medidas de desempeño de este modelo final con el cual posteriormente fueron construidas las métricas. De la tabla anteriormente señalada, se observa una baja *Precision* para el caso de la clase no consumidora, la razón de esto se encuentra en que la mayoría de las personas que contestaron la encuesta había consumido alcohol el último mes.

Clase	***Precision***	***Recall***	**F-Measure**	**ROC Area**
No consumo (0)	0,647	0,918	0,759	0,704
Consumo (1)	0,860	0,590	0,632	0,699
Ponderado	0,754	0,739	0,696	0,701

Tabla 6.13: Rendimiento de *SVM* para la detección de consumo de alcohol en usuarios

Fuente: Elaboración Propia

En la tabla 6.14 se encuentran los pesos normalizados para cada variable, ellos hablan de la importancia que tienen en el modelo. El mayor poder predictivo recae en la variables:

1. Emisión de *tweets* de alcohol
2. Polaridad de políticas
3. Polaridad
4. *Tweets* de consumo
5. Nominaciones externas

[2]En español Máquinas de vectores de soporte

[3]En español Perceptrón multicapa

Las variables 1, 2 y 4 aumentan la probabilidad de consumo de alcohol, mientras que las variables 3 y 5 lo disminuyen.

Por otro lado, las variables más débiles en el modelo son la densidad y la emisión de *tweets* de políticas.

De estos resultados se observa que publicar *tweets* de alcohol (de cualquier tipo) en *Twitter* habla en gran medida de que esa persona consume alcohol.

Como el alcohol es una droga legal en nuestro país, la gente puede hablar de su consumo libremente en las redes sociales sin sentirse perseguida, a diferencia de lo que ocurre con la marihuana en donde la importancia de los *tweets* de consumo juegan un rol menor para ayudar a predecir el consumo de esa droga[18].

Debido al peso de las nominaciones externas también se concluye que el comportamiento del entorno influye en el consumo de alcohol de las personas.

Atributo	**Peso Normalizado**
Edad	-0,98
Tweets de alcohol	1,89
Consumo en *tweets*	1,26
Políticas en *tweets*	0,37
Polaridad	-1,31
Polaridad de políticas	1,84
Seguidores	0,40
Densidad	-0,30
Reach Centrality	-0,97
Uso en vecindario	-0,91
Polaridad en vecindario	1,11
Distancia a consumidores	1,00
Nominaciones externas	-1,12
Tweets de alcohol normalizados por año	1,14
Consumo en *tweets* normalizados por año	1,02
Políticas en *tweets* normalizados por año	0,35
Intercepto	0,022

Tabla 6.14: Influencia de variables en el consumo de alcohol

Fuente: Elaboración Propia

6.4.4. Consumo de riesgo de alcohol en usuarios

La clasificación de consumo de riesgo y dependencia al alcohol presente en los usuarios de *Twitter* fue modelada como un problema de Minería de Datos. Al igual que el modelo de consumo de alcohol en usuarios descrito anteriormente, en esta ocasión no será necesario realizar un tratamiento de textos para conseguir el conjunto de atributos. Para esto fueron utilizados diecisiete atributos, los

cuales están compuestos por medidas calculadas a partir de los *tweets*, el entorno social del usuario, características propias de la persona que posee la cuenta y además, fue incorporado el resultado del modelo anterior, es decir se incorpora la predicción de si el usuario consume alcohol o no.

Este problema puede ser abordado de dos formas, la primera es modelando los puntajes audit como una regresión lineal, ésta fue tomada como primera opción, sin embargo, los resultados obtenidos no fueron satisfactorios. Por esta razón, se prefirió modelar como un problema de clasificación binario, al igual que los modelos anteriores.

Para definir el grupo que presenta consumo de riesgo y dependencia se utilizó el puntaje de corte definido por la OMS, es decir, a partir de los ocho puntos en adelante. Para mayor claridad de las clases se puede observar la tabla 6.15.

Clase	Nombre	**Descripción**
Clase 0	No consumo de riesgo	puntaje AUDIT menor a 8 puntos
Clase 1	Consumo de riesgo	puntaje AUDIT mayor o igual a 8 puntos

Tabla 6.15: Descripción de las clases usadas en modelo de consumo de riesgo

Fuente: Elaboración Propia

Para la construcción de este modelo se utilizaron los algoritmos *Support Vector Machine*, *Multilayer Perceptron* y *Voted Perceptron*. Las medidas de rendimiento de los modelos obtenidos no sufrieron grandes fluctuaciones.

En la tabla 6.16 se muestra las medidas de rendimiento para el modelo final en el cual se utilizó el algoritmo *Support Vector Machine (SVM)*.

En este caso en los datos se observó un predominio de la clase cero, esto afecta al modelo ya que el caso de interés es la clase uno.

Clase	***Precision***	***Recall***	**F-Measure**	**ROC Area**
No consumo de riesgo (0)	0,897	0,550	0.515	0,676
Consumo de riesgo (1)	0,600	0,759	0,738	0,675
Ponderado	0,749	0,660	0,626	0,675

Tabla 6.16: Rendimiento de *SVM* para la detección de consumo de riesgo en usuarios.

Fuente: Elaboración Propia

En este caso no se obtuvo resultados tan buenos, esto se puede explicar por los escasos datos con la etiqueta de clase uno. Por otro lado hay que considerar que en este caso para construir la variable de consumo de alcohol se utilizó el autoreporte que los usuarios entregaron en la encuesta hecha en *Twitter*, esta variable es de tipo binaria, considerando como clase 1 a la gente que ha consumido alcohol el último mes, al igual como se vió en el modelo de consumo de alcohol. Posteriormente en este capítulo se construye la métrica de consumo de riesgo y dependencia para la población general, en esta parte la variable consumo de alcohol se construye con el clasificador de consumo en usuarios.

En la tabla 6.17 se muestran los pesos normalizados para cada una de las variables utilizadas, estos pesos entregan la importancia que tienen al predecir el consumo de riesgo. El mayor poder predictivo lo tienen las siguientes variables:

1. Consumo de alcohol
2. Edad
3. Mención de *tweets* de alcohol
4. Reach Centrality
5. Mención de *tweets* de alcohol ponderado en el tiempo
6. Nominaciones externas

Las variables 1, 3, 4 y 5 aumentan la probabilidad de tener un consumo de riesgo mientras que las variables 2 y 6 lo disminuyen.

Por otro lado, las variables más débiles para predecir el consumo de riesgo o dependencia son la polaridad y el uso en el vecindario.

Se observa que nuevamente juegan un rol importante las variables del entorno social. Publicar *tweets* que hablen de alcohol al igual que en el caso anterior entrega información valiosa del comportamiento de consumo del usuario.

Atributo	**Peso Normalizado**
Consumo de alcohol	1,9468
Edad	-1,9282
Tweets de alcohol	1,2677
Consumo en *tweets*	-0,3328
Políticas en *tweets*	-0,4218
Polaridad	-0,0613
Polaridad de políticas	-0,8266
Seguidores	0,5531
Densidad	0,2211
Reach Centrality	1,2153
Uso en vecindario	-0,1779
Polaridad en vecindario	0,5993
Nominaciones externas	-1,0473
Tweets de alcohol normalizados por año	1,0680
Consumo en *tweets* normalizados por año	-0,7794
Políticas en *tweets* normalizados por año	0,3757
Intercepto	1,7281

Tabla 6.17: Influencia de variables en el consumo de riesgo de alcohol

Fuente: Elaboración Propia

6.5. Métricas

En esta sección se presenta el resultado final de este trabajo. Estas métricas permiten entender el fenómeno del consumo de alcohol situándose en el contexto de *Twitter* a nivel agregado. También es posible extraer algunas conclusiones en un contexto más amplio como lo es para la población general.

6.5.1. Prevalencia

En epidemiología, la prevalencia se define como el porcentaje de la población que evidencia cierta característica dado un periodo de tiempo.

Para llevar a cabo esta métrica, se necesita la recolección de información desde *Twitter*, los clasificadores de texto para ambos casos, políticas y consumo y por último, el clasificador de consumo de alcohol en usuarios.

En este caso el clasificador de consumo en usuarios permite determinar el consumo de alcohol en usuarios el último mes. Si bien en la Encuesta Nacional de Drogas hay tres ventanas temporales; prevalencia en la vida, prevalencia el último año y prevalencia el último mes; solo pudo ser implementada ésta última, ya que por un tema de proporción de los datos arrojados de la encuesta en usuarios, solo se tenía una proporción adecuada para crear los modelos cuando se consideraba la prevalencia en el último mes.

En la figura 6.9 se muestra el cálculo de prevalencia mensual para cada año, a partir desde el 2008 hasta el 2016. El presente año (2017) no fue incluido porque podría verse distorsionado por la cantidad de *tweets* que se poseen.

En la figura 6.10 se muestra el gráfico de la prevalencia mensual por año arrojada por la Encuesta Nacional de Drogas.

Se observa que existe una gran similitud en la tendencia entre la curva predicha por el modelo diseñado, expuesto en la figura 6.9 y la curva publicada en la Encuesta Nacional de Drogas de la figura 6.10. Esta similitud se da comparando los años entre 2008 y 2014 del último gráfico y los años 2010 y 2016 del primer gráfico. Es decir, es necesario retrasar la curva construida en este trabajo dos años y además ponderarla. El desfase podría estar producido por la elección de las variables utilizadas en el modelo. Es necesario aclarar que estos resultados están condicionados por el contexto de *Twitter*, por lo que el análisis es que el comportamiento de la población en general no es reflejado inmediatamente en el contexto de esta red social. Probablemente, primero los usuarios tienen que presentar un comportamiento para luego animarse a generar contenido en relación a él.

En la figura 6.11 se observan las dos curvas graficadas juntas, para construir este gráfico, fue necesario realizar un escalamiento y aplicar el desfase de dos años explicado anteriormente de la curva mostrada en 6.9, es decir, la curva realizada a partir de los algoritmos vistos en este trabajo.

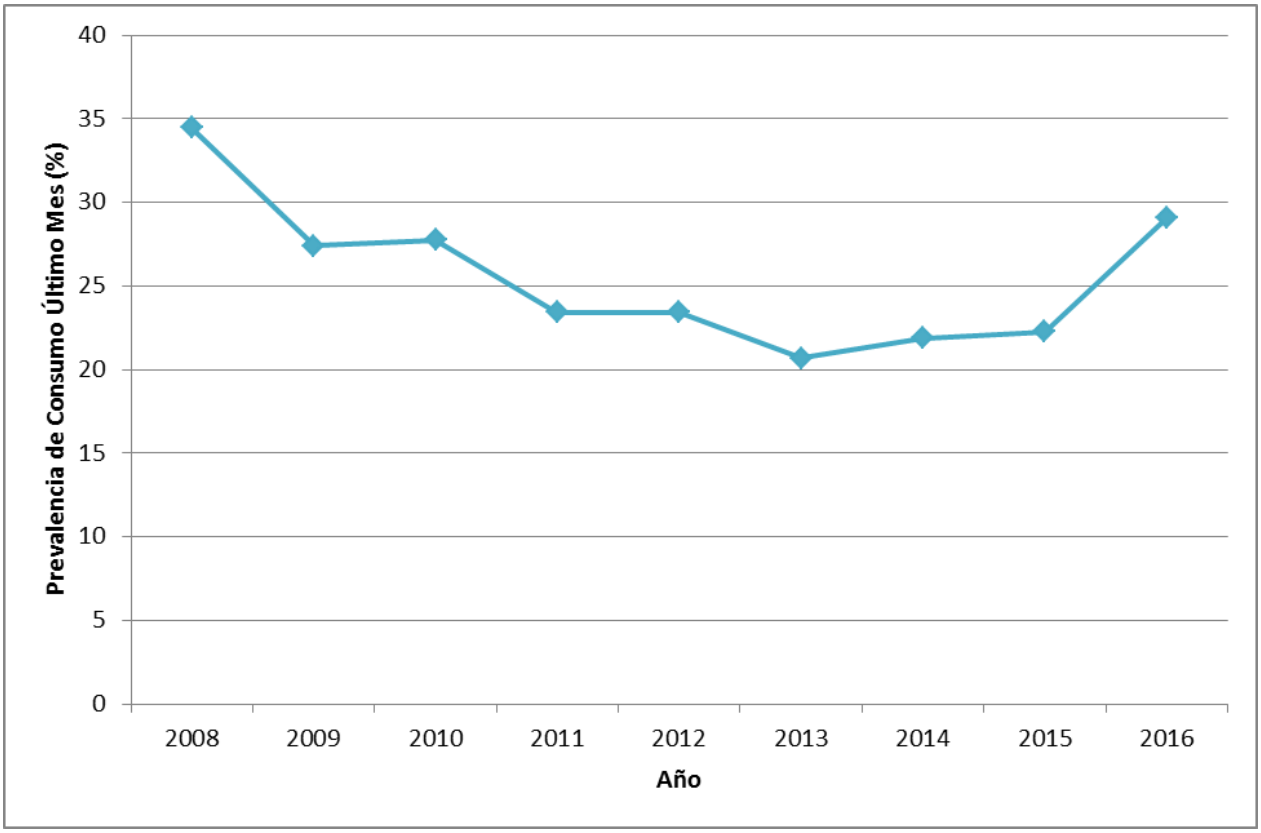

Figura 6.9: Evolución de la prevalencia de consumo de alcohol en el último mes

Fuente: Elaboración propia

6.5.2. Consumo de riesgo

Para llevar a cabo esta métrica se necesitan todos los modelos creados en este trabajo, es decir, la recolección de información desde *Twitter*, los clasificadores de texto para ambos casos, el clasificado de consumo de alcohol en usuarios y el clasificador de consumo de riesgo y dependencia en usuarios.

Como ya se indicó anteriormente se tomó como consumo de riesgo de alcohol la definición tomada por la OMS (Puntaje AUDIT mayor a 8 puntos), esto con el fin de poder comparar los resultados con los que están disponibles en la Encuesta Nacional de Drogas. En la END se presentan dos métricas, una que es la proporción de prevalentes que presentan consumo de riesgo considerando solamente a las personas que han consumido alcohol el último año. La segunda métrica se construye calculando la proporción de prevalentes que presentan consumo de riesgo de alcohol sobre el total de la población. La métrica construida de este trabajo recoge la idea de esta última medida.

La figura 6.12 muestra la evolución de la prevalencia de consumo de riesgo de alcohol y dependencia para cada año sobre el total de la población general. Para construir esta métrica, al igual que el caso anterior, fue tomada una muestra representativa de usuarios para cada año.

En la figura 6.13 se muestra la evolución del consumo de riesgo entregado por la Encuesta

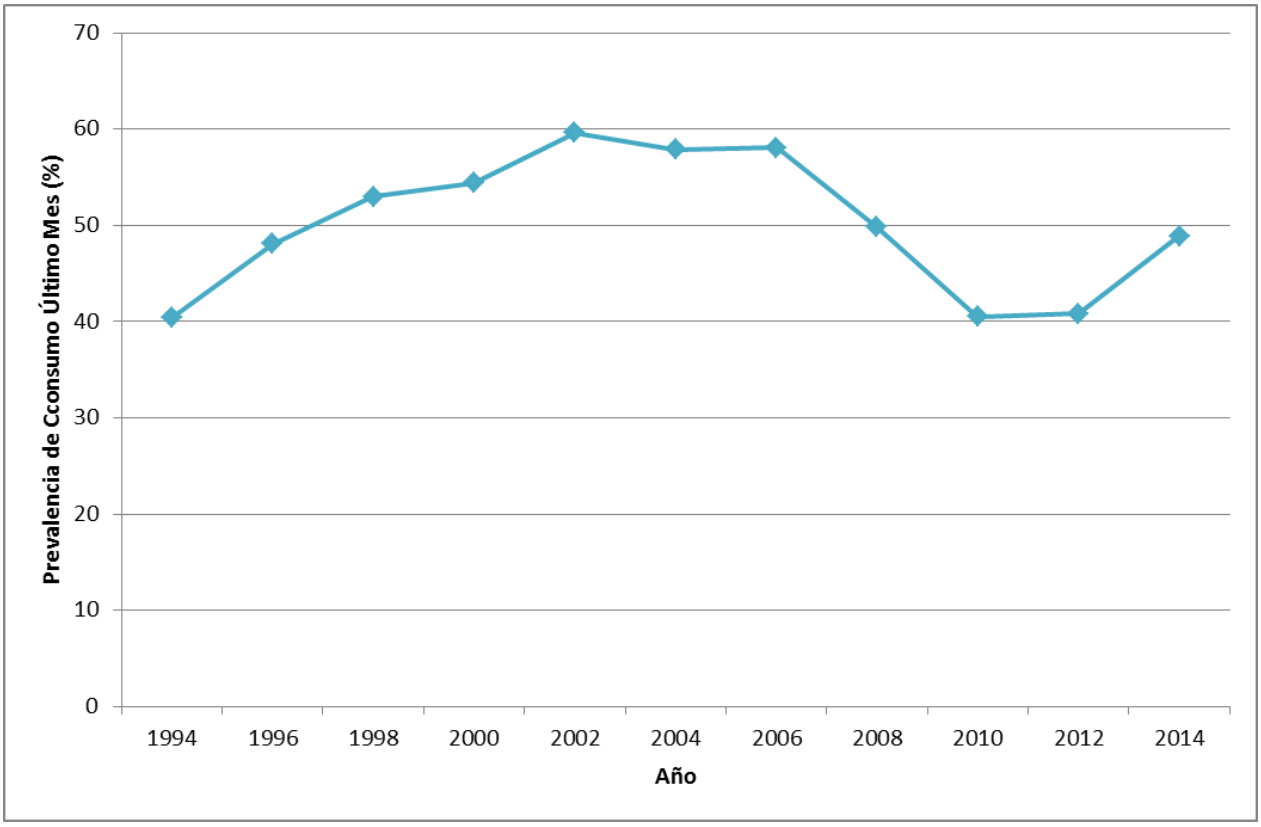

Figura 6.10: Evolución de la prevalencia de consumo de alcohol en el último mes

Fuente: Décimo Primer Estudio Nacional de Drogas en Población General de Chile, 2014

Nacional de Drogas, este es un instrumento que desde el año 2008 el SENDA ha incorporado a su encuesta.

En la figura 6.14 se muestran graficadas juntas las dos curvas, la entregada por el modelo construido en este trabajo y la curva de la Encuesta Nacional de Drogas. Para construir la curva predicha por el modelo se aplicó el desfase de dos años explicado anteriormente.

6.5.3. Frecuencia de consumo

Uno de los desafíos de este trabajo era tener un algoritmo que detectara el consumo de alcohol en los *tweets* emitidos. A partir de este modelo se puede realizar el cálculo de la frecuencia de consumo. En la figura 6.15 se muestra los *tweets* de consumo por año. Este gráfico se obtiene a partir del promedio de los *tweets* de consumo por año por las personas que emiten *tweets* de consumo de alcohol. En esta curva se observa un alza a través de los años.

En la figura 6.16 se observa la frecuencia de uso medida por la Encuesta Nacional de Drogas.

Ambas siguen una tendencia similar, sin embargo es necesario realizar un escalamiento y aplicar el desfase de dos años que se comentó anteriormente.

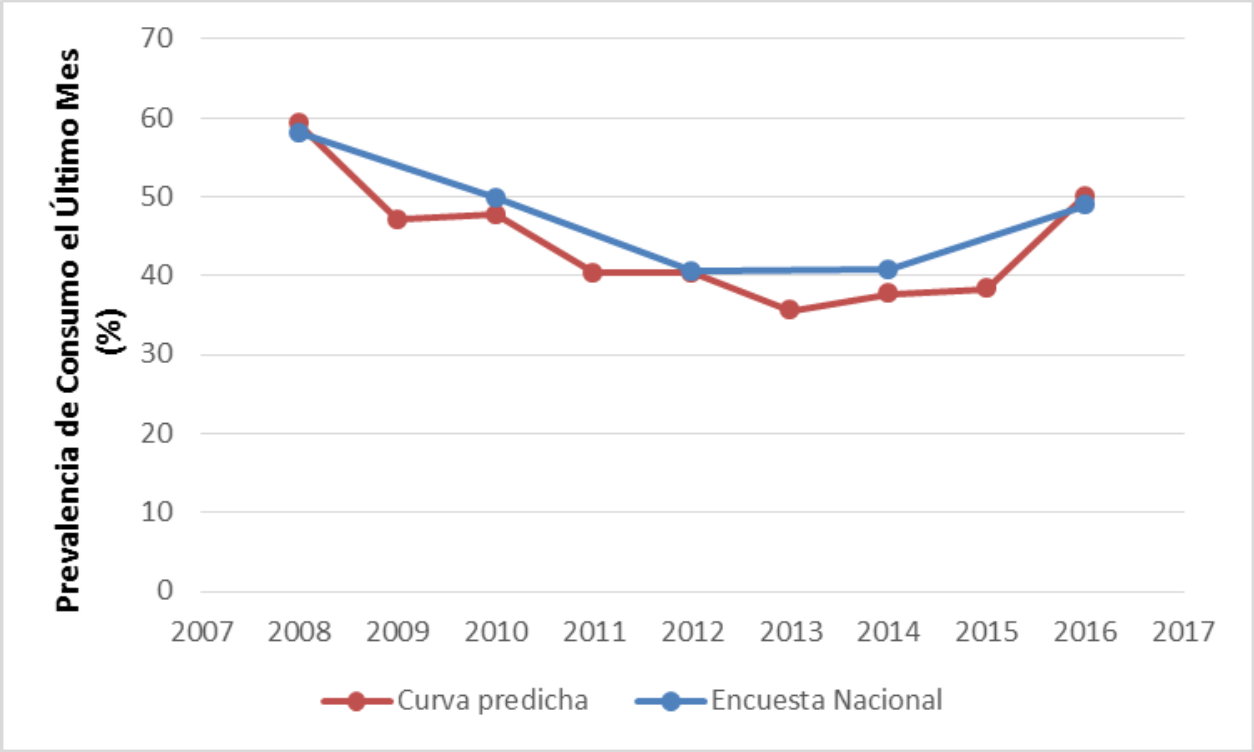

Figura 6.11: Comparación de las dos curvas de evolución de la prevalencia de consumo de alcohol en el último mes.

Fuente: Elaboración Propia

6.5.4. Polaridad

Una de las tareas centrales en el análisis de sentimiento [4] es la clasificación de la polaridad de un texto dado, es decir ver que tan positivo, negativo o neutro es un texto con respecto a una escala. Esta métrica hace un seguimiento de la opinión de las personas para un cierto tema. Se utiliza el promedio de esta métrica para tener un efecto agregado. La figura 6.17 muestra la evolución anual para esta métrica, se consideraron todos los *tweets* de alcohol.

Por otra parte se calculó una polaridad en base a los usuarios. Cada uno de los usuarios tomados para construir esta métricas tiene un número de *tweets* de alcohol que ha emitido y cada uno de esos *tweets* posee una polaridad, lo que resulta en que cada uno de los usuarios tiene una polaridad como resultado de sus propios *tweets*. Como existen usuarios que no escriben sobre alcohol, entonces ellos tendrán polaridad cero. Para calcular esta métrica se tomó el promedio entre los *tweets* del usuario y luego el promedio de la polaridad entre todos los usuarios para cada año. El gráfico de esta métrica se puede apreciar en la figura

[4] *en inglés Sentiment Analysis*

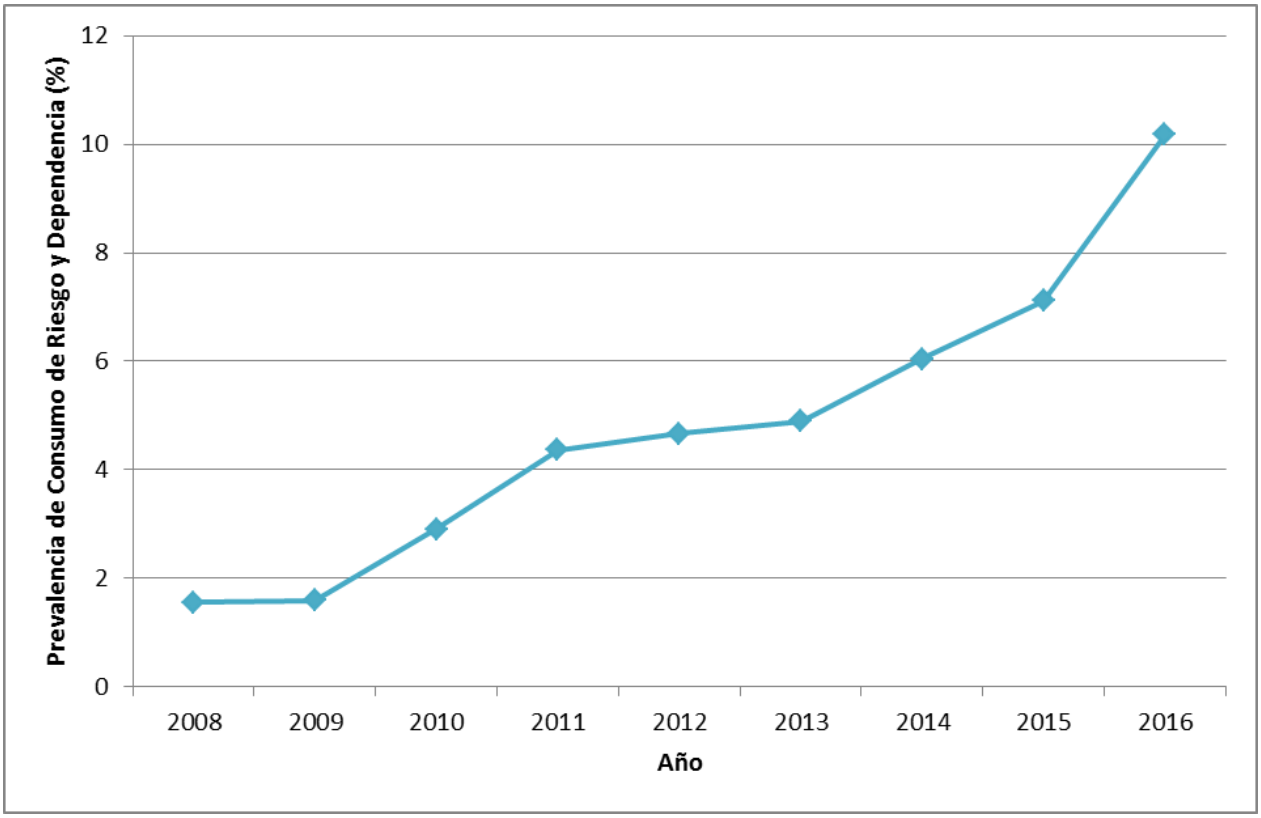

Figura 6.12: Evolución del consumo de riesgo y dependencia sobre el total de la muestra por cada año

Fuente: Elaboración propia

6.5.5. Polaridad de políticas

Para construir esta métrica se siguió el mismo la misma definición dada anteriormente de polaridad, pero la diferencia es que esta vez solo se tomo en consideración los *tweets* clasificados como *tweets* que hacen mención a políticas. La figura 6.19 muestra el comportamiento de esta métrica a los largo de los años.

6.5.6. Palabras más utilizadas

En esta sección se da a conocer las palabras más utilizadas, dependiendo de las categorías de los *tweets*. Este grupo de palabras frecuentemente utilizadas puede dar información respecto a cuales son las bebidas preferidas por la población chilena.

Tweets de consumo

En la tabla 6.18 se muestra en orden descendente los treinta términos con mayor frecuencia en *tweets* de consumo, estas estadísticas fueron construidas en base a los *tweets* que fueron clasificados

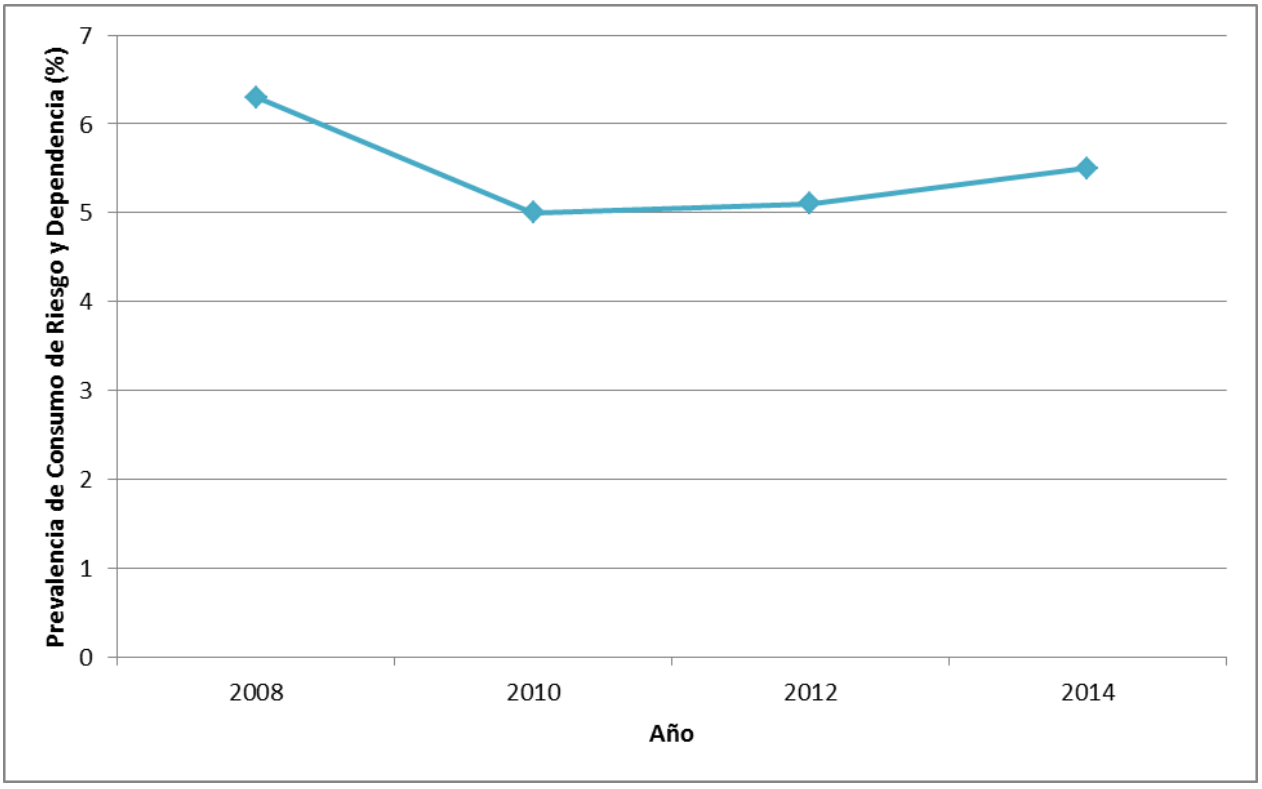

Figura 6.13: Evolución del consumo de riesgo y dependencia sobre el total de la población general

Fuente: Fuente: Décimo Primer Estudio Nacional de Drogas en Población General de Chile, 2014

positivamente en la categoría de consumo.

Se aprecia una amplia variedad de términos, algunos correspondientes al nombre genérico de la bebida alcohólica, otros a los nombres que reciben las preparaciones producto de las combinaciones de alcohol con otro elementos, tales como frutas u otros ingredientes y también están presentes palabras usadas en contextos coloquiales. Destaca que en las palabras se aprecia un fenómeno muy típico presente en el lenguaje chileno que es el hablar con diminutivos.

Los resultados obtenidos se relacionan con las cifras arrojadas por el estudio de *Euromonitor* publicado en el año 2016, las cuales se pueden observar en la tabla 6.19. En ambos casos se muestra una clara preferencia de los consumidores por la cerveza. El hecho de que el vino aparezca en una posición más baja en la tabla 6.18 en comparación al estudio de *Euromonitor*, se explicaría por la edad de los usuarios que integran la red social de *Twitter*, podría influir que el vino es más consumido por personas de rangos etarios mayores, los cuales no posean una cuenta en *Twitter*, o también presenten otro comportamiento, por ejemplo que utilicen su cuenta de *Twitter* para informarse y consumir contenido, pero que no lo generen o no comparten *tweets* acerca de su consumo de alcohol o de sus preferencias.

N°	Palabra	Frecuencia (%)
1	cerveza	17,45
2	pisco	16,45
3	chela	9,51
4	michelada	4,70
5	mojito	3,94
6	vodka	3,92
7	pilsen	3,34
8	copete	3,26
9	trago	3,22
10	alcohol	3,17
11	tequila	3,14
12	chelita	3,06
13	vinito	2,24
14	curao	1,85
15	whisky	1,82
16	tinto	1,73
17	cervecita	1,28
18	terremoto	1,10
19	champagne	1,02
20	ebrio	0,93
21	capel	0,86
22	birra	0,84
23	schop	0,82
24	fernet	0,75
25	ponche	0,68
26	champaña	0,64
27	roncito	0,61
28	espumante	0,59
29	traguito	0,59
30	licor	0,55

Tabla 6.18: Frecuencia de palabras en *tweets* de consumo.

Fuente: Elaboración Propia

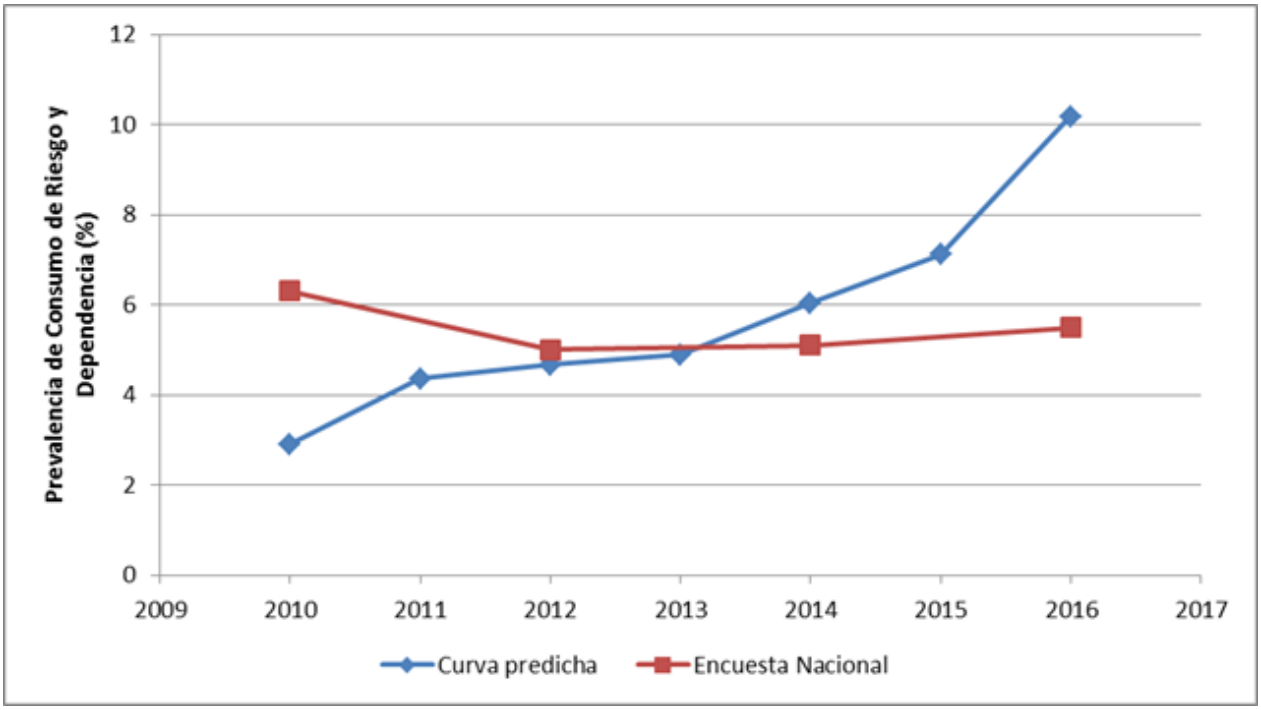

Figura 6.14: Comparación de las dos curvas de evolución de la prevalencia de consumo de riego de alcohol.

Fuente: Elaboración Propia

***Tweets* de políticas.**

En la tabla 6.20 se muestra en orden descendente las veinte palabras con mayor frecuencia en *tweets* relacionados con políticas, estas estadísticas fueron construidas a partir de los *tweets* que fueron clasificados positivamente en la categoría de políticas por los algoritmos construidos.

Se observan algunas diferencias en comparación al caso anterior. En esta ocasión llama la atención que más de la mitad de los *tweets* en los que se menciona alguna política se utiliza la palabra alcohol, la palabra más formal para referirse a esta droga.

Tampoco está presente el uso de diminutivos para referirse a las bebidas alcohólicas.

Por otro lado se aprecia que están presentes algunas palabras que hacen referencia al estado de una persona que ha bebido en exceso, como es el caso de “ebrio”, “curao” y “borrachera”.

A diferencia de la lista de palabras para el caso de consumo, en el caso de *tweets* de políticas la variedad de palabras, es más restringida. En la tabla 6.18 el 90 % de los *tweets* contienen alguna de las veinticuatro primeras palabras. En cambio en la tabla 6.20 para obtener el 90 % de los *tweets* son necesarias diez palabras. Esto ayuda a explicar el rendimiento de los algoritmos de clasificación de *tweets*.

Bebida Alcohólica	Per Cápita anual (litros al año)	Total (en millones de litros)
Cerveza	43,7	784,6
Vino	13,4	240,8
Pisco	1,6	28,6
Cocktail (preparados a base de pisco y vodka)	1,0	17,4
Ron	0,6	10,8
Whiskies	0,5	8,5
Licores	0,1	1,8
Tequila	0,1	1,0

Tabla 6.19: Consumo de bebidas alcohólicas en Chile en el año 2015.

Fuente: Euromonitor

N°	Palabra	Frecuencia (%)
1	alcohol	52,41
2	ebrio	9,63
3	copete	7,49
4	cerveza	5,73
5	pisco	4,41
6	curao	3,90
7	trago	2,60
8	whisky	1,48
9	tinto	1,31
10	capel	1,21
11	licor	1,08
12	chela	0,94
13	schop	0,91
14	vodka	0,82
15	borrachera	0,50
16	tequila	0,48
17	champaña	0,43
18	terremoto	0,42
19	espumante	0,41
20	vino	0,38

Tabla 6.20: Frecuencia de palabras en *tweets* de políticas.

Fuente: Elaboración Propia

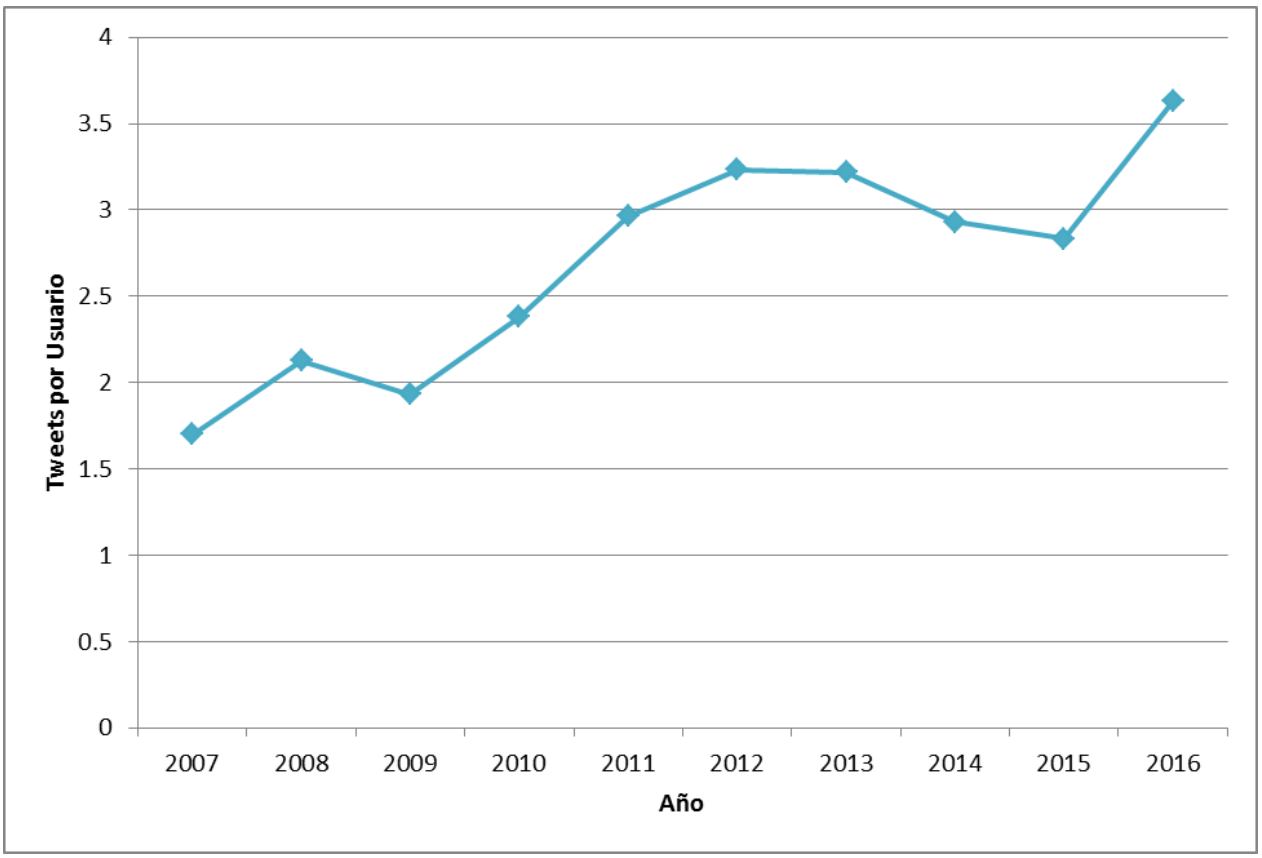

Figura 6.15: Promedio de *Tweets* de Consumo por Usuario por Año

Fuente: Elaboración propia

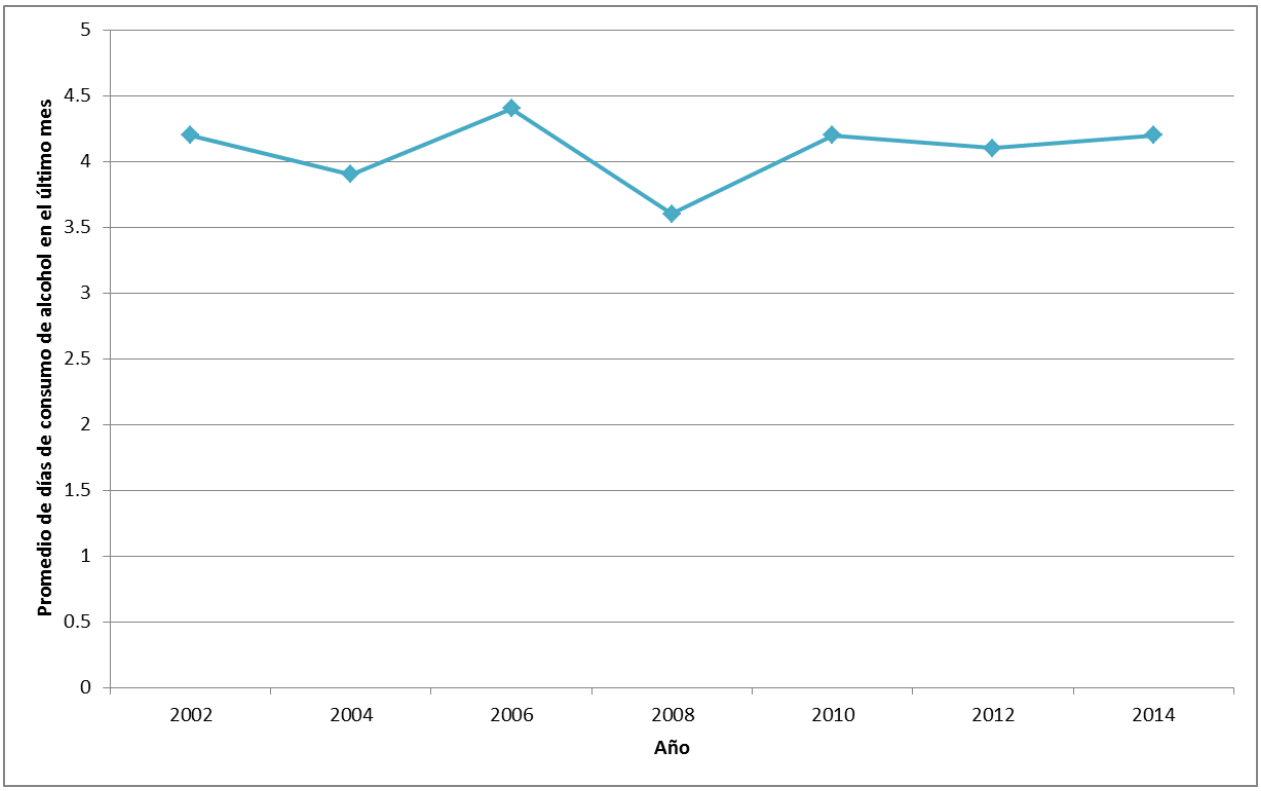

Figura 6.16: Evolución del promedio de días de consumo de alcohol en el último mes.

Fuente: Décimo Primer Estudio Nacional de Drogas en Población General de Chile, 2014

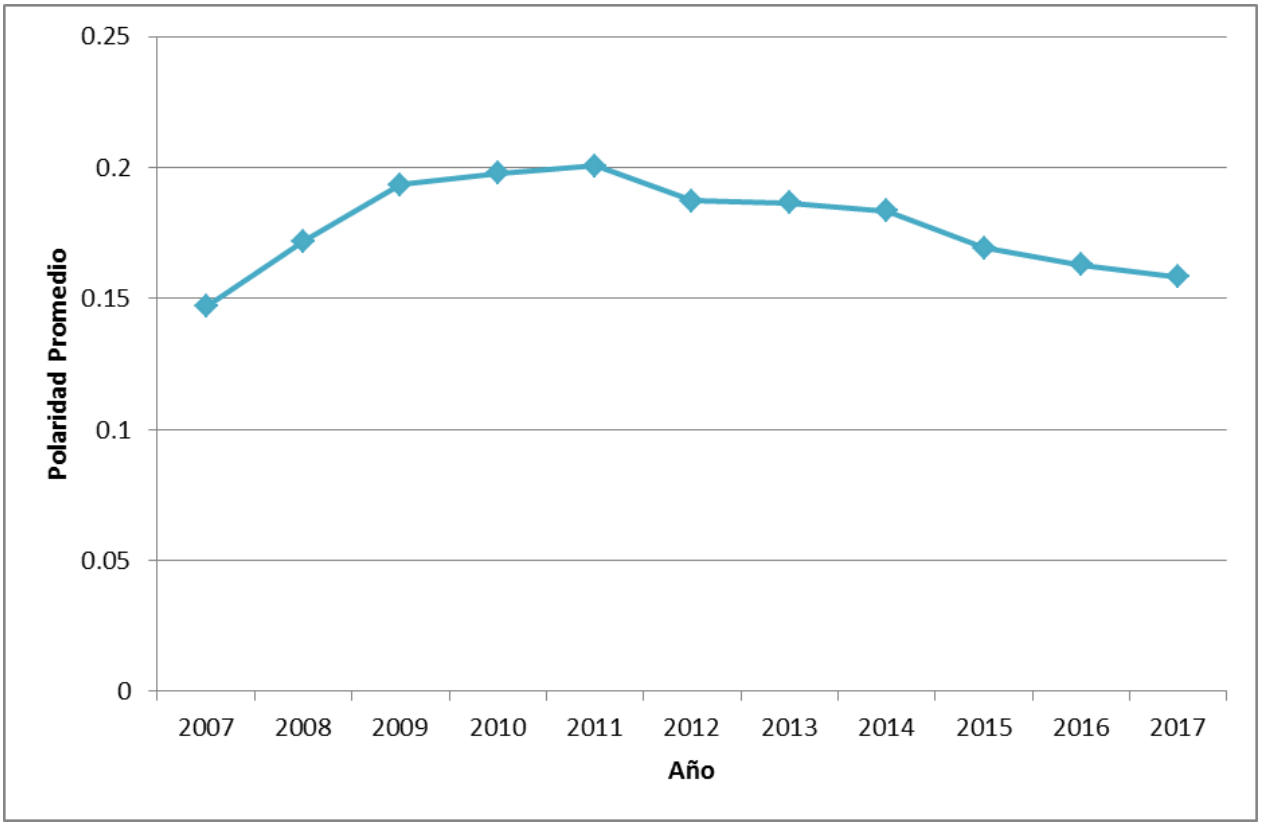

Figura 6.17: Evolución del promedio de la polaridad en los *tweets* de alcohol.

Elaboración Propia

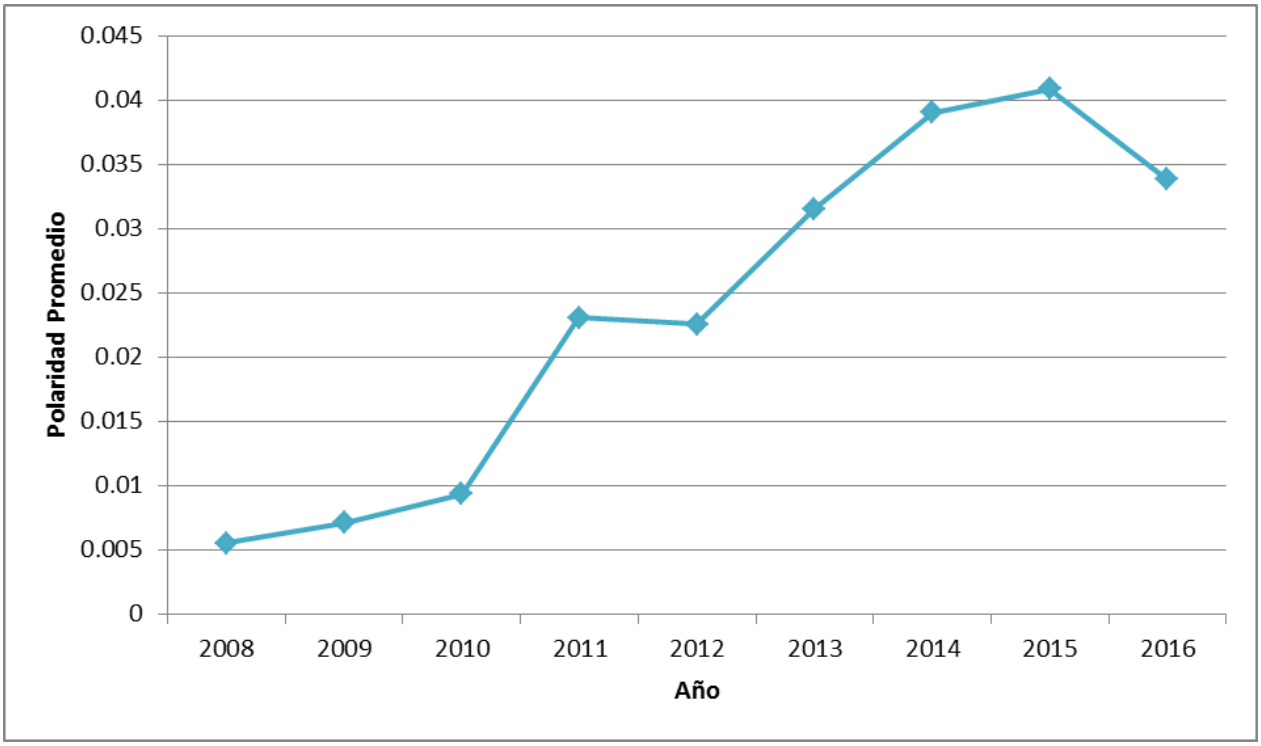

Figura 6.18: Evolución del promedio de la polaridad de los usuarios.

Elaboración Propia

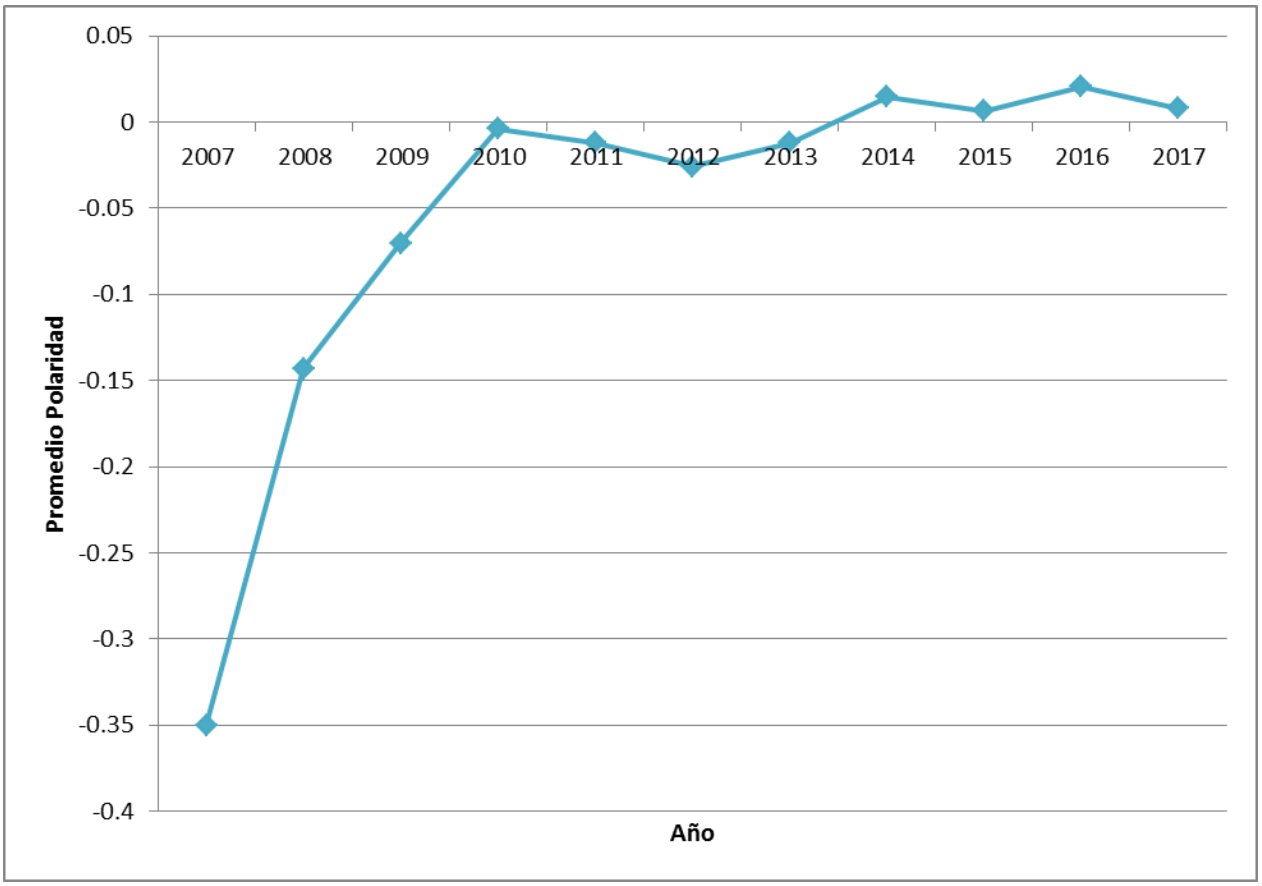

Figura 6.19: Evolución del promedio de la polaridad de los *Tweets* de políticas.

Elaboración Propia

Capítulo 7

Trabajo futuro y conclusiones

Este capítulo final trata sobre los desafíos que se presentan por delante, las mejoras que se pueden llevar a cabo y las conclusiones que se desprenden del trabajo realizado en esta memoria.

7.1. Conclusiones generales

En este trabajo se plantea la utilización de la información disponible en *Twitteer* además del contenido generado por los usuarios con el fin de replicar el comportamiento del consumo de alcohol en la población general.

Uno de los aportes más relevantes de esta aplicación es que permite extraer información valiosa de los *tweets*, los cuales corresponde a textos que tienen la gracia de que en 140 caracteres se condensa todo el mensaje y se presentan como información no estructurada.

El desempeño obtenido para los clasificadores de texto resultó ser satisfactoria. Esto se sustenta en que para la clasificación de *tweets* de consumo de alcohol se obtuvo una *precision* de 0,842 % para la clase de interés y un 0,856 % ponderado, mientras que para la clasificación de *tweets* de políticas se obtuvo una *precision* de 0,977 % para la clase de interés y un 0,944 % ponderado. Sin embargo, fueron apreciadas diferencias con respecto a cada caso de clasificación. El modelo de clasificación de *tweets* de políticas tiene un rendimiento superior al modelo de clasificación de *tweets* de consumo de alcohol. Esto deja en evidencia que dependiendo de tema que se requiera analizar de los *tweets*, se podría necesitar un tratamiento distinto para cada caso particular, ya que algunas veces la división de clases es más difusa debido a que existen distintos matices que hacen un poco más compleja esta separación o se requiere más información de contexto.

En el caso de los clasificadores de usuarios, los resultados obtenidos muestran un desempeño más modesto en comparación a los algoritmos de clasificación de texto. Se puede concluir que para el caso estudiado resulta más difícil modelar y hacer predicciones del comportamiento de los usuarios, probablemente esto se deba a que existen variables que no se pueden rescatar del contexto de *Twitter*.

Una de las grandes ventajas de *Twitter* es que existe un alto porcentaje (91 %) de usuarios que mantiene sus cuentas públicas. Esto fue primordial para poder realizar este trabajo, ya que se pudo recolectar gran cantidad de información, lo que facilitó el aprendizaje de los algoritmos utilizados. Además, las tareas se vieron facilitadas por la API de *Twitter*. Se espera que *Twitter* siga teniendo este espíritu para que continúen los estudios en el área de *Sentiment Analysis*.

Uno de los aprendizajes más relevantes es la importancia del proceso de etiquetado, ya que es una de las tareas centrales para obtener modelos con valores altos de *Recall* y *Precision*. Cuando se trabaja con modelos de clasificación, tener un set de datos que estén etiquetados correctamente, es decir que las personas que revisaron los *tweets* hayan revisado en detalle cada uno de los elementos, es transcendental para posteriormente construir buenos algoritmos que efectivamente logren separar las clases.

Para el caso del consumo de alcohol en usuarios, las variables que juegan los roles más importantes son en primer lugar los *tweets* sobre alcohol que el propio usuario emite y en segundo lugar la polaridad de las políticas. Probablemente esta última se explica por que las políticas acerca de alcohol afectan principalmente a los consumidores, y se traducen en restricciones o un mayor costo al momento de comprar las bebidas alcohólicas. El hecho que el alcohol sea una droga legal repercute en que el usuario se siente libre para expresar su consumo.

Se valoran el uso de relaciones entre usuarios como variables predictivas para el caso de consumo de riesgo y dependencia, en especial las variable *Reach Centrality* y Nominaciones externas.

Finalmente, se concluye que con este trabajo queda demostrado que es factible llevar a cabo lo que fue planteado en la hipótesis de investigación descrita al comienzo, es decir que los algoritmos y modelos desarrollados en la presente memoria permiten extraer información relevante y útil del contenido disponible en *Twitter* para así poder replicar algunas métricas de la Encuesta Nacional de Drogas en materia de consumo de alcohol en la sociedad chilena.

7.1.1. Ética

Es importante destacar que el fin de este estudio es entregar los resultados de manera agregada y no de manera individual para cada usuario. Esto ya que los modelos tienen un porcentaje de error asociado, lo cual se traduce en que dado un usuario el modelo podría clasificarlo de manera incorrecta, es decir, predecir que el usuario tiene un consumo de riesgo cuando en realidad no es así.

Si bien este estudio se plantea como una herramienta para poder entender de mejor manera el consumo de alcohol con el fin de crear iniciativas que eduquen a la población sobre su consumo responsable y sin excesos, alguien podría utilizar estos resultados para fines no éticos como son la discriminación de personas en distintos contextos, por ejemplo en una entrevista de trabajo alguien podría discriminar a un postulante por los resultados entregados por el clasificador de dependencia y consumo de riesgo.

También es importante recalcar que esta aplicación no pretende reemplazar el diagnóstico realizado por un profesional de la salud.

Por esta razón, a pesar que los algoritmos de *Social Media* pueden ser tremendamente útiles, no se debe caer en la utilización de ellos para fines no éticos o que dañen a las personas.

7.2. Trabajo futuro

En esta sección se presentan propuestas para futuras investigaciones que pueden ser incorporadas para pulir el trabajo realizado.

7.2.1. Mejoras a los modelos de clasificación de *Tweets*

Como trabajo a futuro se plantea mejorar los clasificadores de *tweets* para consumo y políticas. Esto se podría realizar etiquetando una mayor cantidad de *tweets* o tener un conjunto de *tweets* etiquetados por un número mayor de personas. Una de las alternativas disponibles sería utilizar herramientas de *crowdsourcing* [1], como lo es por ejemplo *Amazon Mechanical Turk*, o se podrían buscar alternativas adaptadas a Chile o utilizar alguna plataforma con características similares. La idea de estos sitios es que una persona puede hacer trabajos relativamente simples a las que se les paga un pequeño porcentaje por sus tareas realizadas.

Para el caso de *tweets* de consumo de alcohol se podía separar en tres categorías, es decir agregar una categoría más de las que se fueron utilizadas en este trabajo. Se propone, por ejemplo, una primera categoría en la que el *tweet* no hable sobre consumo de alcohol, una segunda categoría en la que se puede deducir el consumo de alcohol por parte del autor con la información que disponible en el *tweet* y una tercera, en la que se tiene sospechas de consumo, pero que no se puede afirmar, es decir agregar una categoría "ambigua".

Incluso se podría agregar más categorías, ya que como se discutió en los capítulos anteriores, los *tweets* de consumo poseen una amplia variedad de situaciones que resulta interesante de analizar. Se proponen a continuación una categoría a modo de ejemplo: una categoría en la que se hable de consumo de alcohol de terceras personas, alguien distinto a quien escribe el *tweet*, quizás esta diferenciación entre los distintos casos posibles ayude al desempeño de los clasificadores.

También resulta interesante estudiar el fenómeno de los *retweets*, analizar si el hecho de realizar un *retweet* implica que se comparte la opinión de quien emitió el *tweet* y si, además, en el caso de los *tweets* que hablan de consumo de alcohol implica el uso de esa droga por parte de quien hace un *retweet*.

Una mejora en las métricas de rendimiento de los modelos de clasificación de *tweets* de consumo y de clasificación de *tweets* de políticas, tendrá un impacto positivo en la precisión de los modelos de prevalencia de consumo de alcohol y en el modelo de consumo de riesgo, ya que este último modelo incorpora como variables a los resultados que se obtienen de los dos primeros modelos.

[1] Este término podría ser traducido al español como colaboración abierta distribuida o externalización abierta de tareas.

Incorporación de *hashtags*

También se puede incorporar el uso de los *hashtag* con los cuales las personas etiquetan sus fotografías con el fin de otros usuarios puedan llegar fácilmente a ellas. Para este estudio se utilizó un filtro para que los *hashatag* fueran eliminados, ya que no es trivial separar las palabras cuando esta tarea está automatizada. Una primera aproximación para poder utilizar la información contenida en los *hashatag* sería utilizar las mayúsculas para reconocer donde empieza cada una de las palabras. Esta sugerencia no resuelve de todo el problema ya que en algunas ocasiones los usuarios escriben la totalidad del *hashatag* con mayúscula o minúscula. La incorporación de esta mejora sería de gran utilidad ya que muchos usuarios utilizan los *hashatag* para entregar información acerca de la bebida alcohólica que están consumiendo en ese momento.

En la tabla 7.1 se muestran algunos ejemplos de *tweets* en los que se puede apreciar que gran parte de la información acerca del consumo se encuentra contenida en el *hashtag*.

Tweet
Ya listos para ver el partido??? ansioso! #CervezasHeladas #CarneAsada #Familia
Justo lo que necesitaba despues de un día de trabajo super pesado.. :) #tierramojada #bar #excelentemúsica #cervezasheladas #nuevagente
Su buena #Pelicula amerita su par de #ChelasHeladas #VolveralFuturo3 https://www.instagram.com/p/BSU7aKHARLL/
Con este sol no se puede estar tranquilo... Solo me da ganas de #beber #cerveza #tengosed #chelasheladas
Matando el calor con el vecino odioso @robgarc #Verano2016 #QueCalor #ChelasHeladas https://www.instagram.com/p/BAiUJN9u3ST/
Su buena siesta accidental, sus buenos huevos a la ostra y ahora #viernesdepiscolas para terminar la semana del terror
#MeGustaCuando te sirvo con tres hielos y te relleno con cocacola #PiscolaTeAmo

Tabla 7.1: *Tweets* en los que la información está contenida en el *hashtag*

Fuente: Elaboración Propia

7.2.2. Mejoras a los modelos de clasificación de usuarios

Modelo de consumo de alcohol

El clasificador de prevalencia de consumo de alcohol en usuarios se podría mejorar, por ejemplo, obteniendo más respuestas a la encuesta de usuarios o incorporando nuevas variables.

Una de las alternativas para obtener un mayor número de respuestas podría ser publicitando de forma masiva la encuesta. Al igual como se llevó acabo en el presente trabajo, hay que tomar

ciertas precauciones debido a lo sensible de la información, por lo que resulta importante contar con el consentimiento informado.

Por otro lado, una de las variables que se propone incorporar está relacionada con ver si el usuario sigue a cuentas que vendan alcohol, o que promuevan el consumo de esta sustancia o cuentas que representan a alguna marca en particular de alcohol, por ejemplo cuentas de marcas de distintos tipos de alcohol o si el usuario interactúa por medio de las menciones de alguna de dichas cuentas. Algunos ejemplos de esas cuentas son: @CervezaCristal, @CampanarioChile o @Closdepirque.

Incorporación del análisis de imágenes

En *Twitter* se observa que muchos usuarios comparten fotografías mientras consumen alguna bebida alcohólica, las cuales generalmente tienen asociado un enlace a *Instagram*. Por esta razón, se podría ampliar el estudio de alcohol a otras redes sociales, como por ejemplo *Instagram*, en la que se necesitaría realizar análisis de fotografías.

Incorporación de emojis relacionados con alcohol

Uno de los problemas de la comunicación en linea a través de las redes sociales es que esta carece del lenguaje no verbal, presente en la comunicación cara a cara. Para suplir esta necesidad se ha popularizado el uso de los emojis, que son ideogramas o caracteres usados en mensajes electrónicos y sitios web.

La Emojipedia[2] es un repositorio que contiene todos los *emojis* disponibles para distintas plataformas, entre ellas *Twitter*. En este sitio se puede encontrar la descripción del *emoji*, como es visualizado en distintas redes sociales, además de su *Codepoints*. En la tabla 7.2 se muestran los *emojis* relacionados con alcohol, la incorporación de estos elementos puede ser incorporada como variable en los modelos de clasificación.

Análisis de los textos contenidos en la descripción del usuario

Algunos usuarios expresan abiertamente que consumen alcohol en el espacio disponible para este realice una descripción de si mismo. Generalmente, en esta sección el usuario describe sus gustos o detalles que son importantes para él.

En la tabla 7.3 se muestran algunos ejemplos de descripciones en donde el mismo usuario reconoce que consume alcohol. Una de las alternativas que se propone es buscar si en la descripción existe alguna de las palabras que hacen referencia al alcohol, pueden ser utilizadas las mismas palabras claves usadas para extraer los *tweets* desde *Twitter* disponibles en la tabla 6.1.

[2]http://emojipedia.org

Emoji	Descripción	Codepoints
	Una copa de vino, que generalmente contiene vino tinto.	U+1F377
	Una cerveza contenida en un vaso o humpen, se muestra con espuma derramándose por el borde superior.	U+1F37A
	Un par de cervezas chocando entre sí en modo de celebración	U+1F37B
	Una botella de champán o vino espumante utilizada comúnmente en celebraciones como por ejemplo la víspera de Año Nuevo.	U+1F37E
	Dos copas de champán que tintinean durante una celebración	U+1F942
	Un vaso de vidrio a menudo asociado con bebidas alcohólicas como el whisky o el ron, se muestra con cubos de hielo y líquido en su interior.	U+1F943
	Una bebida con sabor tropical, que generalmente se sirve con una bombilla, una guarnición (adorno hecho con una rodaja de limón) y un paragüas de cóctel, puede ser referido como una bebida tiki o cóctel con paragüas	U+1F379
	Una copa de cóctel con un tronco delgado y el área superior amplia, a menudo se asocia con los martinis, la mayoría de las versiones de este emoji muestran una aceituna en el interior de la copa.	U+1F378

Tabla 7.2: Tabla de Emojis asociados a alcohol.

Fuente: Imagenes extraída de Emojipedia

Descripción
Me creo periodista en el sitio web de un conocidísimo canal de televisión. Amante de la Católica, la piscola y las giras sin romper nada.
Periodista y Cientista Político. Cruzado a morir. No hay como la piscola.
Periodista, amante de la buena vida, el sour, la piscola y el humor negro. Información es poder. Mamá del niño más lindo del mundo.
22 años, estudio y me gustan las piscolas.
Licenciada en Artes y Humanidades UC, saxofonista, Kiltro lover, piscolera, chelera y devoradora de papas fritas :3
23. Geminiana, tengo un salchihijo, soy #doglover ,cervecera y piscolera. @pablosquash es mi copiloto y amo los atardeceres. ADOPTA NO COMPRES!
Me gusta la piscola cabezona con dos hielos. No me creo más que tú. Más plebeya que princesa. Le rezo a la Santa Sara. Unagi.
Definitivamente sin filtro. Amante de la piscola. No como sal. No tengo cosquillas en lo pies. he intentado q me guste la chela pero no hay caso.

Tabla 7.3: Descripción de usuarios en donde manifiestan que consumen alcohol

Fuente: Elaboración Propia

Modelo de consumo de riesgo

Se propone mejorar el clasificador de consumo de riesgo, buscando otras variables que puedan explicar el comportamiento de las personas que poseen algún grado de dependencia al alcohol y su comportamiento en las redes sociales. Para esto, sería de gran utilidad tener un mayor número de datos para poder generar los modelos de clasificación. Además, en el caso de obtener mejoras en los desempeños del modelo de prevalencia, esto se traduciría en que la variables para representar el modelo de dependencia son más precisas y así, se lograría mejorar las métricas de este último modelo.

Determinación del género de los usuarios

Finalmente, se propone crear un clasificador del género de los usuarios. En *Twitter* esta información no está disponible de manera pública, y tampoco es posible acceder a ella mediante la API, por lo que se necesita un modelo de segmentación.

En el caso de otras drogas como la marihuana, existe evidencia que incluir la variable del sexo del usuario no se traduce en una mejora de los algoritmos de clasificación[18]. Sin embargo, en el caso del alcohol, como ya se ha comentado anteriormente, los hombres tienden a consumir más alcohol que las mujeres, por lo que esta variable podría ayudar a explica de mejor manera el fenómeno del alcohol en *Twitter*.

Bibliografía

[1] E. L. Abel and R. J. Sokol, "Incidence of fetal alcohol syndrome and economic impact of fas-related anomalies," *Drug and alcohol dependence*, vol. 19, no. 1, pp. 51–70, 1987. [Online]. Available: http://dx.doi.org/10.1016/0376-8716(87)90087-1

[2] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of twitter data," in *Proceedings of the Workshop on Languages in Social Media*, ser. LSM '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 30–38. [Online]. Available: http://dl.acm.org/citation.cfm?id=2021109.2021114

[3] Alexa, "twitter.com Traffic Statistics," 2016. [Online]. Available: http://www.alexa.com/siteinfo/twitter.com

[4] M. E. Alvarado, M. L. Garmendia, G. Acuña, R. Santis, and O. Arteaga, "Validez y confiabilidad de la versión chilena del alcohol use disorders identification test (audit)," *Revista médica de Chile*, vol. 137, no. 11, pp. 1463–1468, 2009. [Online]. Available: http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0034-98872009001100008&nrm=iso

[5] C. Aracena Cornejo, "Estudio de la relación entre neurodatos, dilatación pupilar y emocionalidad basado en técnicas de minería de datos," 2014. [Online]. Available: http://repositorio.uchile.cl/handle/2250/115629

[6] T. F. Babor, R. Caetano, S. Casswell, G. Edwards, N. Giesbrecht, and K. Graham, *Alcohol: no ordinary commodity: research and public policy*. Oxford University Press, 2010. [Online]. Available: https://global.oup.com/academic/product/alcohol-no-ordinary-commodity-9780199551149?cc=cl&lang=en&

[7] T. F. Babor, J. C. Higgins-Biddle, J. B. Saunders, and M. G. Monteiro, "The alcohol use disorders identification test. guidelines for use in primary care," *Geneva: World Health Organization*, 2001.

[8] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 20–29, Jun. 2004. [Online]. Available: http://doi.acm.org/10.1145/1007730.1007735

[9] A. Bifet and E. Frank, *Sentiment Knowledge Discovery in Twitter Streaming Data*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 1–15. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-16184-1_1

[10] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[11] T. Brants, "Tnt: A statistical part-of-speech tagger," in *Proceedings of the Sixth Conference on Applied Natural Language Processing*, ser. ANLC '00. Stroudsburg, PA, USA: Association for Computational Linguistics, 2000, pp. 224–231. [Online]. Available: http://dx.doi.org/10.3115/974147.974178

[12] S. Camino, "Desarrollo e implementación de un sistema para identificar tópicos de interés de usuarios chilenos en Twitter," 2016. [Online]. Available: http://repositorio.uchile.cl/handle/2250/144085

[13] N. V. Chawla, *Data Mining for Imbalanced Datasets: An Overview*. Boston, MA: Springer US, 2010, pp. 875–886. [Online]. Available: http://dx.doi.org/10.1007/978-0-387-09823-4_45

[14] J. Cohen, "A coefficient of agreement for nominal scales." pp. 37–46, 1960.

[15] ——, "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit." *Psychological bulletin*, vol. 70, no. 4, pp. 213–220, 1968.

[16] E. Constantinides and S. J. Fountain, "Web 2.0: Conceptual foundations and marketing issues," *Journal of direct, data and digital marketing practice*, vol. 9, no. 3, pp. 231–244, 2008. [Online]. Available: http://link.springer.com/article/10.1057%2Fpalgrave.dddmp.4350098

[17] R. Cooley, B. Mobasher, and J. Srivastava, "Web mining: Information and pattern discovery on the world wide web," in *Tools with Artificial Intelligence, 1997. Proceedings., Ninth IEEE International Conference on*. IEEE, 1997, pp. 558–567.

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

[19] S. de Impuestos Internos, "Indicator Metadata Registry," 2016. [Online]. Available: http://www.sii.cl/aprenda_sobre_impuestos/impuestos/impuestos_indirectos.htm

[20] O. P. de la Salud, *Mortalidad por suicidio en las Américas: informe regional. Washington D.C.* Organización Panamericana de la Salud, 2014.

[21] M. de Transportes y Telecomunicaciones; Subsecretaría de Transportes, "Ley Tolerancia Cero: Modifica Ley Número 18.290, aumentando las sanciones por manejo en estado de ebriedad, bajo la influencia de sustancias estupefacientes o sicotrópicas, y bajo la influencia del alcohol," 2012. [Online]. Available: http://www.leychile.cl/N?i=1037847&f=2012-03-15&p=

[22] ——, "Ley Emilia: Modifica la ley del tránsito, en lo que se refiere al delito de manejo en estado de ebriedad, causando lesiones graves, gravísimas o, con resultado de muerte," 2014. [Online]. Available: http://www.leychile.cl/N?i=1066775&f=2014-09-16&p=

[23] M. del Interior; Subsecretaría del Interior, "Ley Sobre Expendio y Consumo de bebidas alcohólicas," 2003. [Online]. Available: https://www.leychile.cl/Navegar?idNorma=220208

[24] Emojipedia. (22 de Agosto de 2013) `http://emojipedia.org`.

[25] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI magazine*, vol. 17, no. 3, p. 37, 1996.

[26] J. L. Fleiss, "Measuring nominal scale agreement among many raters." *Psychological bulletin*, vol. 76, no. 5, p. 378, 1971.

[27] E. G. Flores, "5 razones por las cuales debes usar PostgreSQL," 2015. [Online]. Available: https://guiadev.com/5-razones-por-las-cuales-debes-usar-postgresql/

[28] C. A. Gonçalves, C. T. Gonçalves, R. Camacho, and E. C. Oliveira, "The impact of pre-processing on the classification of medline documents." in *PRIS*, 2010, pp. 53–61. [Online]. Available: https://web.fe.up.pt/~niadr/PUBLICATIONS/2010/PRIS_2010.pdf

[29] G. Grefenstette, "Tokenization," in *Syntactic Wordclass Tagging*. Springer, 1999, pp. 117–133.

[30] U. Grittner, S. Kuntsche, K. Graham, and K. Bloomfield, "Social inequalities and gender differences in the experience of alcohol-related problems," *Alcohol and Alcoholism*, vol. 47, no. 5, p. 597, 2012. [Online]. Available: +http://dx.doi.org/10.1093/alcalc/ags040

[31] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*. Elsevier, 2011.

[32] J. D. Hawkins, J. W. Graham, E. Maguin, R. Abbott, K. G. Hill, and R. F. Catalano, "Exploring the effects of age of alcohol use initiation and psychosocial risk factors on subsequent alcohol misuse," *Journal of Studies on Alcohol*, no. 3, p. 280, 1997.

[33] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, Sept 2009.

[34] M. A. Hearst, "Text data mining: Issues, techniques, and the relationship to information access," in *Presentation notes for UW/MS workshop on data mining*, 1997, pp. 112–117.

[35] Internet World Stats, "What is Api?" 2016. [Online]. Available: http://www.securitysupervisor.com/security-q-a/computer-security/260-what-is-api

[36] A. Java, X. Song, T. Finin, and B. Tseng, "Why we twitter: understanding microblogging usage and communities," in *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*. ACM, 2007, pp. 56–65.

[37] M. Kobayashi and K. Takeda, "Information retrieval on the web," *ACM Computing Surveys (CSUR)*, vol. 32, no. 2, pp. 144–173, 2000.

[38] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: One-sided selection," in *In Proceedings of the Fourteenth International Conference on Machine Learning*

(ICML), vol. 97, Nashville, USA. Morgan Kaufmann, 1997, pp. 179–186.

[39] S. Kuperman, S. S. Schlosser, J. Lidral, and W. Reich, "Relationship of child psychopathology to parental alcoholism and antisocial personality disorder," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 38, pp. 686–692, 1999.

[40] D. Lester, *Alcoholism and drug abuse. In Assessment and Prediction of Suicide*. Guilford Press, 1992.

[41] F. Leyton and P. Arancibia, "El consumo de alcohol en Chile: Situación epidemiológica." 2016. [Online]. Available: http://www.senda.gob.cl/media/estudios/otrosSENDA/2016_Consumo_Alcohol_Chile.pdf

[42] J. M. López, "Impacto social y económico del abuso del consumo de alcohol," *Economía de la salud*, pp. 82–3, 2005.

[43] L. S. Lowe, "125 Amazing Social Media Statistics You Should Know in 2016," 2016. [Online]. Available: https://socialpilot.co/blog/125-amazing-social-media-statistics-know-2016/

[44] C. Lupton, L. Burd, and R. Harwood, "Cost of fetal alcohol spectrum disorders," *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, vol. 127C, no. 1, pp. 42–50, 2004. [Online]. Available: http://dx.doi.org/10.1002/ajmg.c.30015

[45] O. Maimon and L. Rokach, *The Data Mining and Knowledge Discovery Handbook*. Springer US, 2005, vol. 2. [Online]. Available: http://www.springer.com/us/book/9780387098227

[46] P. Margozzini and J. Sapag, "El consumo riesgoso de alcohol en Chile: tareas pendientes y oportunidades para las políticas públicas," *Tema de la Agenda Pública UC*, vol. 75, 2015.

[47] J. W. Miller, T. S. Naimi, R. D. Brewer, and S. E. Jones, "Binge drinking and associated health risk behaviors among high school students," *Pediatrics*, vol. 119, no. 1, pp. 76–85, 2007.

[48] O. J. Morgan and C. H. Lizke, *Family Interventions in Substance Abuse: Current Best Practices*. Routledge, 2013.

[49] X.-H. P. . C.-T. Nguyen, "JGibbLDA," 2008. [Online]. Available: http://jgibblda.sourceforge.net/

[50] M. Olavarría, "Estudio nacional sobre costos humanos, sociales y económicos de las drogas en chile, 2006," *Santiago, Chile, Olavarría y Asociados*, 2008. [Online]. Available: http://www.senda.gob.cl/media/estudios/otrosSENDA/Costos%20Humanos%20Sociales%20y%20Economic%20Drogas%20en%20Chile%202008.pdf

[51] W. H. Organization, "Estrategia mundial para reducir el uso nocivo de alcohol," 2010. [Online]. Available: http://www.who.int/substance_abuse/activities/msbalcstrategyes.pdf

[52] ——, *Global status report on alcohol and health*. World Health Organization, 2014. [Online]. Available: http://apps.who.int/iris/bitstream/10665/112736/1/9789240692763_eng.pdf

[53] ——, "Indicator code book, global information system on alcohol and health," 2014. [Online]. Available: http://www.who.int/substance_abuse/activities/gisah_indicatorbook.pdf?ua=1

[54] ——, *Preventing suicide: a global imperative*. World Health Organization, 2014.

[55] ——, "Indicator Metadata Registry," 2017. [Online]. Available: http://apps.who.int/gho/data/node.wrapper.imr?x-id=2376

[56] L. Padró, "A hybrid environment for syntax-semantic tagging," *CoRR*, vol. cmp-lg/9802002, 1998. [Online]. Available: http://arxiv.org/abs/cmp-lg/9802002

[57] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008. [Online]. Available: http://dx.doi.org/10.1561/1500000011

[58] S. Popova, S. Lange, L. Burd, A. E. Chudley, S. K. Clarren, and J. Rehm, "Cost of fetal alcohol spectrum disorder diagnosis in canada," *PloS one*, vol. 8, no. 4, p. e60434, 2013.

[59] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[60] PostgreSQL, "Hot Standby," 2016. [Online]. Available: https://wiki.postgresql.org/wiki/Hot_Standby

[61] V. R. Preedy and R. R. Watson, *Comprehensive handbook of alcohol related pathology*. Academic Press, 2004.

[62] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-validation," in *Encyclopedia of database systems*. Springer, 2009, pp. 532–538.

[63] B. Ridout, A. Campbell, and L. Ellis, "'off your face(book)': Alcohol in online social identity construction and its relation to problem drinking in university students," *Drug and Alcohol Review*, vol. 31, no. 1, pp. 20–26, 2012. [Online]. Available: http://dx.doi.org/10.1111/j.1465-3362.2010.00277.x

[64] J. Schulenberg, J. L. Maggs, T. E. Dielman, S. L. Leech, D. D. Kloska, J. T. Shope, and V. B. Laetz, "On peer influences to get drunk: A panel study of young adolescents," *Merrill-Palmer Quarterly*, vol. 45, no. 1, pp. 108–142, 1999. [Online]. Available: http://www.jstor.org/stable/23093317

[65] *Décimo Primer Estudio Nacional de Drogas en Población General de Chile, 2014*, SENDA, Ministerio del Interior y Seguridad Pública, Gobierno de Chile, 2015.

[66] K. J. Sher, K. S. Walitzer, P. K. Wood, and E. E. Brent, "Characteristics of chiledren of alcoholics: Putative risk factors, substance use and abuse, and psychopathology," *Journal of Abnormal Psychology*, vol. 100, pp. 427–448, 1991.

[67] L. Sher, "Alcohol consumption and suicide," *QJM: An International Journal of Medicine*, vol. 99, no. 1, pp. 57–61, 2005. [Online]. Available: http://dx.doi.org/10.1093/qjmed/hci146

[68] V. Shkapenyuk and T. Suel, "Design and implementation of a high-performance distributed web crawler," in *Data Engineering, 2002. Proceedings. 18th International Conference on*. IEEE, 2002, pp. 357–368.

[69] A.-H. Tan, "Text mining: The state of the art and the challenges," in *In Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, vol. 8, 1999, pp. 65–70. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/citations;jsessionid=0B619439F3DCFE18B510EEF024FBC1CE?doi=10.1.1.132.6973

[70] O. Tsur and A. Rappoport, "What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities," in *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM, 2012, pp. 643–652.

[71] Twitter, 2016. [Online]. Available: https://dev.twitter.com/streaming/overview

[72] ——, "About Twitter," 2016. [Online]. Available: https://about.twitter.com/company

[73] ——, "About verified accounts," 2016. [Online]. Available: https://support.twitter.com/articles/119135

[74] J. D. Velásquez and L. C. Jain, *Advanced techniques in web intelligence*. Springer, 2010, vol. 311.