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

DISEÑO E IMPLEMENTACIÓN DE UN SISTEMA PARA MONITOREAR EL CONSUMO Y OPINIÓN SOBRE LA MARIHUANA EN TWITTER

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

VÍCTOR DAVID CORTÉS SÁNCHEZ

PROFESOR GUÍA:
JUAN DOMINGO VELÁSQUEZ SILVA

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

SANTIAGO DE CHILE
2016

RESUMEN DE LA MEMORIA PARA OPTAR AL
TITULO DE: Ingeniero Civil Industrial
POR: Víctor David Cortés Sánchez
FECHA: 08/06/2016
PROFESOR GUIA: Juan Domingo Velásquez Silva

DISEÑO E IMPLEMENTACIÓN DE UN SISTEMA PARA MONITOREAR EL CONSUMO Y OPINIÓN SOBRE LA MARIHUANA EN TWITTER

Este trabajo tiene como objetivo diseñar e implementar una aplicación que recolecte información de los usuarios chilenos de *Twitter* para monitorear el consumo y opinión sobre la marihuana dentro del mismo contexto, y evaluar los resultados con respecto a los valores reales de la población nacional.

La aplicación se sostiene como rama de investigación del proyecto CORFO, "OpinionZoom". El cual está enfocado en explorar grandes bases de datos generadas gratuitamente para recopilar, organizar y extraer conocimiento. Es por esto que fue contactado por la Unidad de Adicciones del Hospital Clínico de la Universidad de Chile para aplicar este enfoque en el estudio de drogas en Chile. Especialmente en la marihuana, cuyo consumo ha evidenciado un crecimiento promedio sostenido durante los últimos años, aumentando los costos asociados a la droga. Por esta razón, se buscan nuevas herramientas que puedan explicar el comportamiento reciente.

La cantidad total de información digital ha explotado en los últimos años, siendo conformada en su mayoría por datos no estructurados. Esto se explica por la mayor participación de los usuarios de sitios web en la creación de contenido. Particularmente, Twitter brinda un ambiente donde pueden compartir libremente, lo cual genera gran cantidad de información relacionada con la vida de sus usuarios.

La aplicación de Text Mining, Data Mining y Web Opinion Mining habilita la extracción de patrones desde datos estructurados y no estructurados para obtener información relevante que apoye la toma de decisiones. La clasificación de textos y los sentimientos emitidos por ellos pueden ser combinados con la estructura de las relaciones entre usuarios para replicar el alto poder predictivo del contorno social con respecto al consumo de marihuana.

La implementación de la aplicación fue realizada en código Java, utilizando el paradigma de programación modular. La aplicación permite extraer *tweets* relacionados con marihuana, clasificarlos con respecto a categorías, extraerles la polaridad y combinarlos con medidas de Análisis de Redes Sociales para predecir el consumo de marihuana. Los resultados señalaron que la combinación de modelos con rendimientos medianamente buenos es útil para predecir el consumo de marihuana a nivel individual. A nivel agregado se obtuvieron resultados prometedoras, pero aún faltan datos para la validación estadística, dejando los resultados a la interpretación del cliente. Se concluye que la información generada en Twitter representa una herramienta poderosa para comprender el comportamiento de las personas y ayudar a la toma de decisiones del estado con respecto a políticas públicas.

A mis padres, mis hermanas y mi perrito, Apolo.

Agradecimientos

Estoy agradecido de la vida por todo lo que me ha brindado. Por sobre todo, agradezco la familia que tengo. No puedo imaginar una vida sin mis padres y mis hermanas. No puedo imaginar una vida sin su cariño y apoyo incondicional. Agradezco de corazón a mis padres por las posibilidades que me han regalado. Si no fuera por ellos, no estaría escribiendo esto.

Pensándolo bien, antes que agradecer a las personas que me han apoyado durante la carrera, agradezco a la carrera y la facultad por haberme permitido conocer a aquellas personas. La facultad es la razón que me impulsó a viajar lejos de mi hogar y descartar cualquier otra opción. Fue el medio para conocer a todos los amigos que me han acompañado y en parte, han remplazado a mi familiar mientras estoy lejos de casa. Gracias a todos por haber estado allí.

Gracias a todos los que pusieron de su parte para que entrara a la universidad. En especial, a mis profesores de matemáticas de educación básica y media por lograr que me interesara en la materia. Gracias al Taekwondo y a mi maestro por lograr que un niño inquieto encontrara la forma de tranquilizarse.

Finalmente, gracias al centro por darme la oportunidad de trabajar en lo que me gusta. Gracias a todas las personas que me acompañaron durante el desarrollo de la memoria. Hicieron que prefiera pasar mi tiempo en la salita, convirtiéndola en mi segundo hogar.

Tabla de Contenido

Índice de tablas

Índice de figuras

Capítulo 1

Introducción

1.1. Contexto de Trabajo

El tema de memoria está enmarcado dentro un proyecto CORFO: *"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 consignadas gratuitamente en la red con el fin de recopilar, organizar y extraer conocimiento. Dicho conocimiento debe ser capaz de ayudar a la toma de decisiones por parte de los organismos interesados.

El tema fue originado desde la Unidad de Adicciones de la Clínica Psiquiátrica de la Universidad de Chile, donde el jefe a cargo es el Doctor Carlos Ibáñez. La Unidad de Adicciones está interesada en disminuir el número de chilenos adictos a las drogas lícitas e ilícitas. Dicho interés es compartido a nivel país, ya que es congruente con los Objetivos Sanitarios 2011-2020 del Ministerio de Salud [23]. La alarma es detonada por las altas tasas de prevalencia evidenciadas a nivel nacional en los últimos años. Se cree que la expansión de esta rama de investigación puede aportar al entendimiento de los consumidores de drogas y subsecuentemente, a la prevención y disminución del consumo.

1.1.1. Situación Actual y Oportunidad

A lo largo de la historia, el consumo de drogas ha sido asociado con efectos negativos en la vida de las personas. La consecuencia más directa es que provoca dependencia en algún nivel, independientemente de la droga, y a su vez, esta dependencia produce otro tipo de problemas relacionados con la calidad de vida y pérdidas para la sociedad. Esto lo convierte en objetivo de estudio e intervenciones. Los cuales están acompañados de grandes esfuerzos y desembolso de dinero por parte del estado [21] y organismos privados.

La marihuana es un caso especial de estudio. Esta droga tiene particular atención a nivel país, debido al amplio debate y relevancia que se le ha dado en Chile y el mundo. En los últimos años, la percepción de riesgo de la droga ha disminuido, reflejando un norma social a favor del consumo

y pudiendo ser origen del alza significativa del consumo nacional (transversal en la población). La tasa anual de prevalencia fue de 7,1 % para el año 2012, una de las más altas a nivel Latinoamericano en ese periodo. Cifra que aumentó aún más en el año 2014 (11,3 %). Por lo tanto, las preocupaciones son justificadas.

La alta prevalencia del consumo de marihuana y el riesgo que continúe aumentando en los próximos años, se percibe como un problema desde los organismos interesados, donde el accionar se concreta en la ejecución de políticas de prevención del consumo, y ejecución de políticas en materia de tratamiento, rehabilitación e integración. Un punto de especial de atención radica en detectar posibles focos de intervención, sobre todo en adolescentes y jóvenes. En efecto, para el segmento de 12 a 18 años y 19 a 25 años la tasa de prevalencia anual del año 2012 era de 6.7 % y 17.5 %, respectivamente. Incluso se observa en [72], que en el año 2014 el primer segmento obtuvo una cifra que asciende a 30.6 %.

Desde los encargados del diseño y monitoreo de políticas públicas con respecto a la marihuana, surge la necesidad de hacer seguimiento al consumo agregado de la droga y mejorar la comprensión acerca de los mecanismos que incentivan tal comportamiento. Por esta razón, el Servicio Nacional para la Prevención y Rehabilitación del Consumo de Drogas y Alcohol (SENDA) realiza estudios para recolectar información sobre la magnitud del consumo, percepción de riesgo, entre otras variables. Esta información es desagregada según características socio-demográficas y socio-económicas.

El costo del Estudio Nacional de Drogas en sus versiones de población general y escolar (MM$ 427,6 y MM$ 171,5, respectivamente) hacen que el periodo entre estudios (cada dos años) sea mayor de lo recomendado. La frecuencia de los estudios dificulta el seguimiento continuo de la evolución de la prevalencia e impide la detección de cambios abruptos de manera temprana. Además no captura variables específicas que podrían explicar de mejor manera los niveles observados de consumo. Por lo tanto, la oportunidad se centra en la creación de fuentes complementarias, que puedan mejorar y enriquecer la calidad de información. Además de aumentar la frecuencia de recolección de datos.

Los costos socio-económicos de la política de drogas en Chile para los años 2008 y 2009 fueron calculados en [21] y representan el 0,61 % y 0,59 % del PIB (MMM$ 574,12 y MMM$ 568,87), respectivamente. En base a esa estimación de costos y a una regresión lineal de la tasa de prevalencia del consumo de marihuana (Figura 1.1), se realizó una estimación aproximada de los costos en función de dicha tasa. Se observa una proyección ascendente de costos (MMM$ $682,58$ para el año 2016), pues se estima que la prevalencia será igual a $8,8$ % en ese año. Los grandes costos que enfrenta la sociedad en este tema ameritan explorar nuevas formas de recolección de datos para mejorar el diseño y aplicación de políticas públicas con respecto al consumo de drogas.

1.1.2. Descripción y Justificación del Proyecto

Actualmente el enfoque de las empresas gira en torno al cliente. Esto significa que buscan comprender a sus clientes para generarles mayor valor a través de sus productos o servicios, y de esta manera capturar mayor beneficio neto. El escenario ideal es segmentar por comportamientos, es decir, realizar todos los esfuerzos de promoción sólo en aquellos individuos que son más propensos

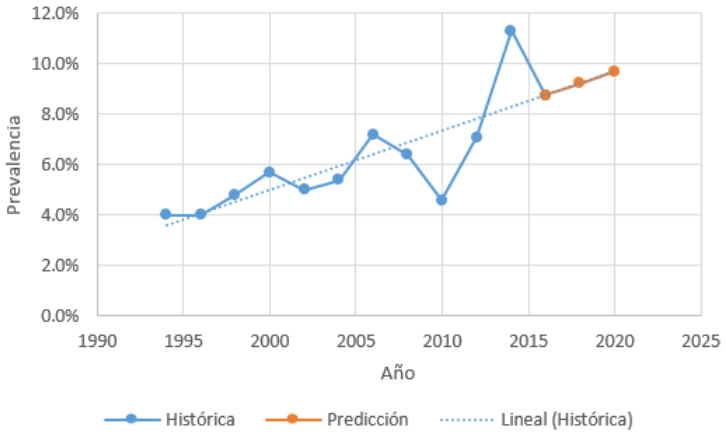

Figura 1.1: Regresión lineal de prevalencia
Fuente: Elaboración Propia

a pagar por el producto o servicio. En particular, se quiere reconocer aquellos factores que identifican a la persona y la condicionan a convertirse en cliente, y cómo estos factores evolucionan a través del tiempo.

La forma en que la empresa busca identificar y caracterizar a sus clientes puede generalizarse y adaptarse a objetivos sociales, en otras palabras, buscar el beneficio social en vez de maximizar beneficios económicos privados. En el caso de las drogas, el Ministerio de Salud, por medio del SENDA, busca disminuir un comportamiento que considera nocivo para la sociedad. Por esta razón, se quiere medir la proporción de individuos que evidencia el comportamiento e identificar los factores individuales y de entorno que influyen en su adopción. En este contexto, la marihuana es reconocida como la droga ilícita más utilizada en la población chilena y es un ejemplo claro de lo anterior. Idealmente se busca identificar a los individuos más propensos a usarla y realizar medidas de prevención para disminuir su uso, sobre todo en adolescentes y jóvenes, ya que es el grupo más afectado.

Cada día, gran cantidad de información es generada mediante plataformas donde los usuarios asumen el poder de creación de contenido. El usuario da a conocer datos personales, intereses, actividades, relaciones e interacciones con otros usuarios. Las personas usan estos escenarios para expresarse y comportarse de manera natural, y destinan gran parte de su tiempo sumergidos en estos ambientes. Por esta razón, estos escenarios son interesantes desde el punto de vista de recolección de información que pueda representar a las personas y sus comportamientos.

En este contexto, la aplicación de técnicas de minado de datos pueden traducirse en ventajas significativas al momento de procesar información sin estructura y en forma de textos, para brindar datos complementarios a los mecanismos usados actualmente. La dificultad de este tipo de iniciativas reside en estructurar gran cantidad de textos presentes en bases de datos y transformarlos en información que ayude a la toma de decisiones.

Debido a que la información está concebida en forma de textos, el proyecto consiste en minar en la gran cantidad de documentos en busca de aquellos que estén relacionados con la marihuana y más precisamente, identificar aquellos textos que impliquen el consumo de marihuana. Donde

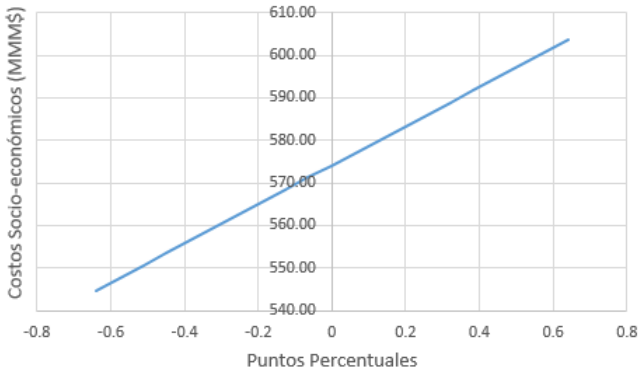

Figura 1.2: Sensibilidad de costos (prevalencia)
Fuente: Elaboración Propia

cada documento está asociado a un usuario en particular haciendo uso de la redes sociales. A partir de esta propiedad es posible calcular métricas que ayuden al estado caracterizar el fenómeno y a tomar decisiones al respecto. Este trabajo se verá materializado en una aplicación web que permita visualizar las métricas antes mencionadas.

Una manera de asociar los beneficios de un estudio como este es realizando una estimación aproximada de la sensibilidad de los costos socio-económicos en función de la tasa de prevalencia del consumo de marihuana. Los supuestos para realizar esta estimación están basados en [21], donde se evaluó un escenario de despenalización de marihuana. Otro supuesto fuerte de la medición consiste en que las variaciones de costos dependen proporcionalmente a la tasa de prevalencia.

Los cálculos aproximados sugieren que una disminución de 0,1 puntos porcentuales de la prevalencia se podría traducir en una disminución de 0,8 % (MMM$ 4,6) de los costos socio-económicos asociados a drogas (Figura 1.2). Esto implica que si las decisiones tomadas a partir de este estudio logran disminuir la prevalencia, los costos socio-económicos podrían decrecer notablemente.

1.2. Objetivos

1.2.1. Objetivo General

Diseñar e implementar un sistema que recolecte datos consignados por los usuarios chilenos de *Twitter* y permita hacer seguimiento al consumo y opinión sobre la marihuana

1.2.2. Objetivos Específicos

- Diseñar y aplicar una metodología y algoritmos que permitan la extracción de contenido estructurado y no estructurado, generado por los usuarios de las redes sociales. Específicamente textos relacionados con el tópico marihuana.
- Utilizar técnicas de Text Mining y Data Mining para extraer patrones de los datos de *Twitter*,

y definir procesos de elaboración de métricas de monitoreo del consumo y opinión sobre la marihuana

- Aplicar técnicas estadísticas para medir la relación entre las métricas construidas y sus respectivas métricas inspiradoras de la Encuesta Nacional de Drogas.
- Diseñar una aplicación que permita el monitoreo continuo de las métricas.

1.3. Hipótesis de Investigación

El trabajo de investigación pretende poner a prueba la siguiente hipótesis:

"Es posible extraer y procesar información de Twitter para representar un fenómeno complejo como el consumo y opinión sobre la marihuana de la población chilena"

1.4. Metodología

El primer paso consiste en realizar una revisión bibliográfica de las dos áreas abarcadas en este trabajo. La primer área estriba en los factores de riesgo y protectores del uso de la marihuana. Esto servirá para conocer la actualidad del tema e identificar oportunidades de desarrollo. La segunda área engloba técnicas de *Text Mining*, *Opinion Mining* y *Data Mining* que ayuden a extraer información sobre el consumo y opinión de la marihuana en las redes sociales.

Luego, se construirán los algoritmos que permitan recolectar los documentos desde las bases de datos de *Twitter*. Estos algoritmos deben ser adecuados para adaptarse a las limitaciones de la fuente y ser capaces de obtener los datos que cumplan con los requerimientos del estudio. Primero se hará uso de bases de datos del Proyecto *OpinionZoom* para refinar el mecanismo de clasificación de documentos (*Topic Model*) y luego se aplicará a la fuente principal para obtener los datos históricos.

A partir de los datos recolectados en el paso anterior, se alimentará a un conjunto de algoritmos de clasificación y regresión. Estos algoritmos permitirán reconocer parámetros en los datos y transformar estos datos brutos en información relevante. Dicha información ayudará a mejorar el entendimiento del consumo y opinión sobre la marihuana y su evolución en *Twitter*.

Se construirán métricas que permita hacer seguimiento al consumo y opinión sobre la marihuana. Dichas métricas serán originadas tomando como modelo las variables incorporadas en el Estudio Nacional de Drogas que cumplan con factibilidad de desarrollo y requerimientos de la Unidad de Adicciones. Subsecuentemente, las métricas serán sintetizadas de tal manera que permitan probar su relación con su respectiva variable origen. Todas estas variables están consignadas en los Estudios Nacionales de Drogas en Población General de Chile, elaborados por el SENDA.

Finalmente, se diseñará una herramienta que permita hacer un seguimiento continuo a la evolución de las métricas sobre el consumo y opinión. La aplicación debe ser consistente con la lista de requerimientos funcionales y visuales establecidos por el cliente. Para este paso se deberá hacer una revisión de las herramientas tecnológicas que permitan cumplir satisfactoriamente con los

objetivos del trabajo.

1.5. Resultados Esperados

El desarrollo del trabajo pretende obtener los siguientes resultados:

- Una metodología que posibilite la extracción de contenido de la web relacionado con un tópico específico y bajo ciertos requerimientos especiales.
- Una base de datos actualizable que almacene contenido relacionado con la marihuana y las métricas construidas.
- Modelos predictores que posibiliten obtener información relevante con respecto al consumo de marihuana en *Twitter*.
- Datos estadísticos que permitan probar la relación entre las métricas caracterizadoras del uso de marihuana y su variable origen.
- Una aplicación que permita extraer contenido relacionado con la marihuana, aplicar procesamiento y sintetizar los datos para la mejor comprensión del cliente.

Capítulo 2

Marco Conceptual

Dada la naturaleza del trabajo propuesto, es necesario explicitar algunos conceptos generales con respecto al tema y dar una reseña de los métodos y técnicas que serán utilizadas.

2.1. World Wide Web

La World Wide Web (W3) es un sistema de información colaborativa heterogéneamente distribuido en forma de documentos. La World Wide Web es concebida como un mundo continuo en donde toda la información puede ser alojada (servidores) y accedida (clientes) desde cualquier fuente.

El principio de W3 de lectura universal es que una vez la información está disponible, debería ser accesible desde cualquier tipo de computadora, en cualquier país, y una persona debería usar sólo un simple programa, para acceder a ella [88]. La red se sostiene en algunos conceptos esenciales:

- Hipertexto: Es un texto con enlace que pueden conducir desde todo o parte de un documento hacia todo o parte de otro documento. Los documentos pueden ser texto, gráficos, vídeos y sonidos (hiper-media).
- Búsqueda: Su lógica se basa en buscar información en grandes cantidades de datos. Para hacerlo, se muestra un panel de búsqueda que permite introducir texto y se obtiene una respuesta de hipertextos que conduce a documentos relacionados con la búsqueda.

La W3 se basa en un conjunto de prácticas comunes que permiten a clientes y servidores comunicarse:

- Uniform Resource Locator (URL): Permite describir cualquier objeto en cualquier parte de internet. Todo documento tiene esta dirección para ser referenciado.
- Hypertext Transfer Protocol (HTTP): Es un protocolo de comunicación que utilizan los elementos de arquitectura web. Basado en un sistema de peticiones y respuestas.
- Hypertext Markup Language (HTML): Es un formato básico que todo cliente W3 entiende. Describe la estructura lógica del documento.

2.2. Web 2.0

El término es definido en [17] como "una colección de código abierto, interactiva y aplicaciones en línea controladas por el usuario que expanden las experiencias, conocimiento y poder de mercado de los usuarios como participantes en negocios y procesos sociales. Las aplicaciones de Web 2.0 soportan la creación de redes de usuarios informales, facilitando el flujo de ideas y conocimiento por medio de la generación eficiente, diseminación, intercambio y edición/refinado del contenido informacional".

Aunque los términos Social Media y Web 2.0 son usualmente asociados al mismo significado, algunos asocian el término Web 2.0 con las aplicaciones online y el término Social Media con los aspectos sociales de las aplicaciones Web 2.0. Cualquier categoría de las aplicaciones Web 2.0 tiene al usuario como factor vital, principalmente como contribuidor de contenido, no sólo como consumidor. Para resumir este atributo es usado el término User-Generated Content (UGC). En [17] se propone una clasificación de aplicaciones Web 2.0 basado en su tipo:

- Blogs: Son diarios online, donde el autor publica de manera periódica secciones referentes a cualquier índole. A su vez, los visitantes pueden comentar, pudiendo generar un diálogo con el autor.
- Redes Sociales: Permiten a los usuarios construir sitios web personales accesibles por otros usuarios. Se facilita el intercambio de información personal y la comunicación entre usuarios. Facebook es el ejemplo más popular.
- Comunidades (contenido): Son sitios web que organizan y comparten contenido de cierto tipo. Un ejemplo de contenido audiovisual es YouTube.
- Foros: Son sitios donde se comparten ideas e información que se relacionan con intereses particulares. Ejemplo de esto es Asshai.com, un foro dedicado a la serie Juego de Tronos.
- Agregadores de contenido: Son sitios web que permiten a los usuarios personalizar completamente el contenido que se desea acceder. Por ejemplo, Google.

En resumen, la gran innovación desde la llamada Web 1.0 hacia la Web 2.0 es la conversión del usuario desde consumidor hacia generador de contenido, donde la interconexión ha aumentado sustancialmente la cooperación e interacción entre los usuarios.

2.3. Extracción de datos

Este paso consiste en extraer una colección de datos que eventualmente será minada bajo el objeto de extraer información relevante. Tal como se menciona en [73], actualmente existen dos enfoques para obtener información desde bases de datos de aplicaciones web. El primer enfoque corresponde al uso de Application Programming Interface (API) propios de cada aplicación web y el segundo consiste en el uso de Web Crawlers.

2.3.1. Application Programming Interfaces (API's)

Las API's son interfaces implementadas por las aplicaciones web que facilitan el acceso a sus bases de datos. Dentro de las API's más famosas se encuentra la Twitter API y la Graph API de Facebook. Generalmente los datos son facilitados en forma de objetos tales como comentarios, usuarios, entidades, entre otros y a través de métodos de consulta que exigen identificación. Comúnmente, el número de consultas es limitado.

Debido a que son específicos a cada proveedor, las API's son fáciles de implementar, pero están restringidas a un ambiente. Los datos obtenidos están estructurados, a excepción de los textos, que por definición son contenido sin estructura. Además no todas las aplicaciones web utilizan este enfoque y están sujetas a cambios a lo largo de tiempo.

A modo de ejemplo, la API de *Twitter* proporciona dos versiones: Streaming API y Rest API. La primera permite obtener los "tweets" a medida que son creados y la segunda, realiza una búsqueda histórica de datos. Para cada modalidad es posible especificar varios parámetros como lenguaje, ubicación, palabras relacionadas, etc.

2.3.2. Web Crawlers

Son algoritmos cuya función es recorrer la Web en búsqueda de información, aprovechando su estructura son capaces de recolectar gran cantidad de información sin más límite que la capacidad de almacenamiento y la funcionalidad misma del algoritmo. Tal como describe [39], la implementación de este enfoque es compleja y los datos obtenidos son semi-estructurados. Este método de adquisición es dividido en dos categorías: crawling incremental y crawling focalizado. La primera consiste en acceder a la mayor cantidad posible de páginas de Internet, ya que la recolección aumenta progresivamente mediante el desempeño del algoritmo y pocas restricciones a la búsqueda. La segunda categoría trata de recolectar datos relevantes para algún tópico en específico, en efecto, la estructura de recorrido es evaluada utilizando listas priorizadas.

2.4. Data Mining y Web Mining

Knowledge Discovery in Databases (KDD), frecuentemente llamado Data Mining se define en [35] como "el proceso de descubrimiento de nuevas correlaciones significativas, patrones y tendencias mediante la examinación de grandes cantidades de datos almacenados en repositorios, usando tecnologías de reconocimiento de patrones así como técnicas estadísticas y matemáticas". Las raíces de Data Mining son trazadas sobre tres líneas: estadísticas clásicas, Artificial Intelligence (AI) y Machine Learning.

Machine Learning (ML) es la unión entre estadísticas y Artificial Intelligence. El intento de ML es permitir que los programas computacionales aprendan sobre los datos que están estudiando, usando estadísticas como conceptos fundamentales y agregando heurísticas y algoritmos de AI para modelar los objetivos. En muchas formas Data Mining se puede entender como la adaptación de

las técnicas de ML hacia aplicaciones de negocios.

ML tiene como objetivo el cumplimiento de dos tareas [56]

- Construir un modelo conciso que represente las relaciones entre atributos.
- Emplear las relaciones de los atributos para predecir datos desconocidos.

Se destacan dos categorías de algoritmos de ML:

- Aprendizaje Supervisado: se intenta inferir una función sobre un conjunto de datos etiquetados. Cada instancia está asociada con un resultado esperado (variable dependiente).
- Aprendizaje no supervisado: en este caso, se pretende que la función describa la estructura escondida de los datos, ya que no están asociados a una etiqueta conocida de antemano.

Web Mining, aunque es considerado una aplicación particular de Data Mining, justifica un campo separado de investigación, principalmente por las características de los datos que analiza. Web Mining puede ser ampliamente definido como el descubrimiento y análisis de información útil desde W3 [67]. Los tipos de datos difieren ampliamente en su contenido (pueden ser texto, imágenes, audio, etc.) y meta-información. Además los datos web se caracterizan por no ser etiquetados, distribuidos, heterogéneos, semi-estructurados, dinámicos y de alta dimensionalidad. Por lo tanto, Web Mining básicamente se enfrenta con gran cantidad de información con hiper-vínculos y dichas características.

Web Mining puede clasificarse en tres tipos:

- Web Structure Mining (WSM): Consiste en minar la estructura de hiper-vínculos dentro de la misma red. La estructura representa el grafo de los enlaces en el sitio y entre sitios.
- Web Usage Mining (WUM): Son usados los datos secundarios generados por la interacción del usuario con la red. WUM trabaja con perfiles de usuarios, patrones de acceso de los usuarios, y mina caminos de navegación.
- Web Content Mining (WCM): Consiste en el descubrimiento de información útil desde contenido de la red. Estos no son sólo textos, ya que abarcan otro tipo de datos como audio, video, meta datos y datos hiper-vinculados. Las técnicas actuales están centradas en el análisis del texto y los hiper-vínculos.

2.5. Text Mining

Text Mining es un campo que intenta obtener información significativa desde textos de lenguaje natural. Comparado con el tipo de datos almacenado en bases de datos, el texto no es estructurado, es amorfo y es algorítmicamente difícil de tratar. En Data Mining, el objetivo es extraer información escondida, desconocida y que puede ser arduamente extraída. Sin embargo, en Text Mining, la información es clara y explícitamente escrita en los textos, la única forma que pueda ser desconocida es la incapacidad de las personas de leerla por sí mismos. Por lo tanto, el objetivo es poder procesar grandes cantidades de información automáticamente, sin intermediación humana.

Según [1] los textos pueden ser analizados en diferentes niveles de abstracción. Los datos pueden ser tratados como una bolsa de palabras donde cada palabra no guarda ninguna relación con otras palabras y pueden ser tratados como una serie de palabras donde el orden de ellas aporta información semántica significativa. Aunque es preferible utilizar el segundo nivel de abstracción, la mayoría de los enfoques utilizan el primero por su mayor simplicidad, sobre todo desde un punto de vista algorítmico y por falta de desarrollo en métodos de Natural Language Processing.

El rango de técnicas de Text Mining es amplio, variando desde Information Retrieval hasta Sentiment Analysis, pero en este apartado se detallarán dos tipos de algoritmos: Unsupervised y Supervised Learning. Esta separación fue mencionada anteriormente, pero el tratamiento de texto hace necesario profundizar las definiciones, ya que es un caso particular de Data Mining. Los enfoques de Topic Modelling y Sentiment Analysis se detallarán de forma independiente.

2.5.1. Unsupervised Learning

En [1] los métodos de Unsupervised Learning son descritos como métodos que no requieren datos de entrenamiento, pudiendo ser aplicados a textos sin ningún tipo de esfuerzo manual. Clustering y Topic Modelling son los dos métodos más usados en el marco de análisis textual. La diferencia entre estos dos métodos radica en que la asignación a categorías o tópicos de cada texto es más cruda en el primero, es decir, cada texto puede pertenecer a un solo tópico. Mientras que el segundo método implementa una asignación más suave, calculando probabilidades de pertenencia a cada tópico.

2.5.2. Supervised Learning

También conocidos como métodos de clasificación por su uso extendido en esta aplicación. Los métodos de Supervised Learning explotan los datos de entrenamiento, es decir, un conjunto de puntos y su correspondiente valor objetivo. Una vez extraídas las relaciones intrínsecas entre los datos, son aplicadas a nuevos datos sin etiquetar o predecir. Estos métodos son ampliamente utilizados en Data Mining y se han ido adaptando para ser utilizados en Text Mining. Las técnicas más populares son los árboles de decisión, Support Vector Machine, Naive Bayes Classifier, entre otras.

2.6. Topic Modelling

Según [7], Probabilistic Topic Modelling es una serie de algoritmos desarrollados por investigadores de Machine Learning con el fin de calificar una gran cantidad de textos con información temática. Tal como se mencionó anteriormente, los algoritmos de Topic Modelling no requieren etiquetamiento previo, facilitando el trabajo del investigador.

El modelo de tópicos más simple es Latent Dirichlet Allocation (LDA), el modelo incorpora la intuición que los documentos pueden contener varios tópicos. Un tópico se define formalmente

como una distribución sobre un vocabulario fijo. Bajo esta definición, en cada documento de una colección de documentos, las palabras son generadas en dos etapas:

- Elegir aleatoriamente una distribución sobre tópicos.
- Para cada palabra en el documento, primero elegir aleatoriamente un tópico desde la distribución de tópicos y luego, elegir aleatoriamente una palabra desde la distribución de palabras para cada tópico.

Desde el punto de vista de descubrimiento de información, la colección de documentos es observada, mientras que los tópicos, la distribución de tópicos por documentos y la asignación de tópicos por palabra en cada documento es una estructura oculta (estructura de tópicos). De esta forma, el problema computacional consiste en inferir esta estructura oculta. Esta dificultad es superada con modelamiento probabilístico, se utiliza la distribución conjunta para calcular la distribución de las variables ocultas dadas las variables observadas. La distribución conjunta llamada distribución posterior se muestra a continuación:

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{\mathrm{i}=1}^{K} p(\beta_{\mathrm{i}}) \prod_{\mathrm{d}=1}^{D} p(\theta_{\mathrm{d}}) (\prod_{n=1}^{N} p(z_{\mathrm{d},n}|\theta_{\mathrm{d}}) p(w_{\mathrm{d},n}|\beta_{1:K}, z_{\mathrm{d},n}))$$

Donde:

- Los tópicos son $\beta_{1:K}$, donde cada β_k es una distribución sobre el vocabulario.
- La distribución de tópicos para el d-ésimo documento son θ_{d}, donde $\theta_{\mathrm{d},k}$ es la proporción para el tópico k en el documento d.
- La asignación de tópicos para el d-ésimo documento son z_{d}, donde $z_{\mathrm{d},n}$ es la asignación de tópico para la n-ésima palabra del documento d.
- Las palabras observadas para el documento d son w_{d}, donde $w_{\mathrm{d},n}$ es la n-ésima palabra del documento d, donde es un elemento desde el vocabulario fijo.

El cálculo computacional de esta ecuación tiene ciertas complicaciones que la hace inmanejable, por lo tanto es necesario hacer algunas aproximaciones, existen varias opciones para esto. Además existen varias expansiones del modelo LDA, que derivan de la relajación de algunos supuestos, en [7] se mencionan algunas.

2.7. Sentiment Analysis

Sentiment Analysis, también llamado Opinion Mining es definido en [60] como el área de estudio que analiza las opiniones, sentimientos, evaluaciones, valoración, actitudes y emociones de las personas hacia entidades tales como productos, servicios, organizaciones, individuos, materias, eventos, tópicos, y sus atributos. Sentiment Analysis se enfoca principalmente en opiniones que expresan o implican sentimientos positivos o negativos.

Debido al crecimiento explosivo de las aplicaciones Web 2.0, individuos y organizaciones han incrementado el uso de este medio para apoyar la toma de decisiones. En [60] se sostiene que

herramientas como estudios, encuestas, y focus groups pueden dejar de ser necesarias para las organizaciones, ya que Opinion Mining permite recolectar opiniones públicas desde un amplio rango de intereses. De la misma manera, puede convertirse en una herramienta complementaria de aquellas mencionadas anteriormente.

2.7.1. Planteamiento del Problema

Con el objetivo de dar una estructura a textos en lenguaje natural no estructurado y establecer las sub-partes del problema a resolver, en [60] se establece la definición más completa y actualizada del problema de Opinion Mining. En esta definición se consideran cinco elementos importantes por cada opinión, conformando una tupla de esos 5 elementos, los cuales son: el objetivo de la opinión, el atributo del objetivo al cual la opinión hace alusión, el emisor de la opinión, el tiempo cuando la opinión fue emitida y la polaridad contenida en la opinión (positiva, neutra o negativa). A continuación se presenta la definición formal del problema en forma de tupla:

$$(\mathrm{e_i}, a_{\mathrm{i}j}, s_{\mathrm{i}jkl}, h_k, t_l)$$

Donde:

- $\mathrm{e_i}$ es el nombre de la entidad i.
- $a_{\mathrm{i}j}$ es el atributo j de la entidad i.
- $s_{\mathrm{i}jkl}$ es sentimiento (polaridad) sobre el aspecto j de la entidad i del emisor k en el tiempo l.
- h_k es el emisor k.
- t_l es el tiempo l cuando la opinión fue emitida por k.

El objetivo principal de Opinion Mining es identificar el conjunto de tuplas para un conjunto de documentos. Cada elemento en la tupla conlleva un sub-problema a resolver, al igual que aglomerar sus soluciones. Para esto, existen varias alternativas que son nombradas en [60].

2.7.2. Diferentes Niveles

En general, el planteamiento del problema de Sentiment Analysis ha sido separado en tres niveles de granularidad:

- Nivel de documento: en este nivel la tarea está en determinar si un documento como un todo expresa sentimientos positivos o negativos. Se asume que un documentos contiene opiniones acerca de una sola entidad.
- Nivel de oración: en este nivel se explora y se determina la polaridad de sentimientos para cada oración en un documento. Se utilizan técnicas para granular el documento y separarlo en una serie de oraciones para analizarlas individualmente. Análogamente al nivel anterior, asume que cada cada oración contiene una sola entidad o aspecto.

- Nivel de aspecto o entidad: es el nivel de análisis más fino de todos. En este nivel se busca directamente cada opinión, en base al supuesto de que cada opinión contiene un sentimiento (positivo o negativo) y su objetivo. El objetivo de una opinión puede ser una entidad en sí misma o sus diferentes atributos (aspectos).

2.7.3. Diferentes Enfoques

Generalmente existen dos grandes enfoques que han sido utilizados para desarrollar la tarea de Opinion Mining. En [6] se nombran estos dos: Unsupervised Lexicon-based y supervised Machine Learning. Además un enfoque relativamente nuevo es Concept-based Opinion Mining. Cada uno de los enfoques tiene características diferentes que permiten explotar la naturaleza subjetiva de los documentos.

- Unsupervised lexicon-based approach: este enfoque explota reglas y heurísticas obtenidas del lenguaje para realizar la tarea de determinar la polaridad de los textos. La idea principal es hacer uso de un diccionario de palabras para determinar en el texto la polaridad de palabras que usualmente se utilizan para reflejar sentimientos (por ejemplo, adjetivos). Luego se utilizan reglas del lenguaje que pueden cambiar o intensificar la polaridad de conjuntos de palabras.
- Supervised Machine Learning: este enfoque consiste en algoritmos que permiten extraer patrones adyacentes en un conjunto de datos clasificados o etiquetados con un valor, llamado conjunto de entrenamiento. Los patrones son obtenidos como relaciones entre variables que representan a los textos y la variable buscada (etiqueta). Comúnmente el objetivo es clasificar o predecir el valor de nuevos datos no etiquetados, es decir, calificar textos según su polaridad o predecir valores que reflejan polaridad de sentimientos.
- Concept-based Opinion Mining: en este enfoque se utilizan modelos que conceptualizan el conocimiento de una materia dada para que sea comprensible por la computadora. Frecuentemente consisten en grafos, donde los nodos son conceptos y las aristas representan las relaciones entre ellos. Estos modelos, también llamados ontologías, se incorporan a otras técnicas para mejorar el proceso de clasificación.

2.8. Teoría de Grafos

Un grafo consiste en un conjunto de nodos y aristas que conectan pares de nodos, por lo general modelan conjuntos de objetos que tienen conexión. En el caso de Twitter, los nodos representan usuarios y las aristas relaciones de amigos y seguidores. Formalmente un grafo está formado por un par $(V(G), E(G))$, siendo $V(G)$ un conjunto de vértices y $E(G)$ un conjunto de aristas que se denotan (i, j). Dos vértices son adyacente si $(\mathrm{i}, j) \in E(G)$, existe un conjunto que abarca a todos los nodos adyacente a i, se llama vecindario de i y se denota como $N_G(\mathrm{i})$. El grado de un vértice es la cardinalidad del conjunto $N_G(\mathrm{i})$ y se escribe como $\mathrm{d}_G(\mathrm{i})$.

2.8.1. Representación Matricial

Se utilizan matrices para modelar conceptos y calcular cifras especiales relacionadas con el área.

- Matriz de Adyacencia: Es la matriz utilizada para representar el grafo, que explica cuáles son los nodos que están conectados entre sí y el tipo de conexión. La matriz de adyacencia se define como:

$$A_{\mathrm{i}j} = \begin{cases} 1 & \text{si } (\mathrm{i}, j) \in E \\ 0 & \text{si no} \end{cases}$$

- Matriz de Incidencia: Grafo de dimensiones $|V|x|E|$ que se basa en el grafo $G = (V, E)$.

$$b_G = \begin{cases} 1 & \text{si } (\mathrm{i}, j) \in \mathrm{e}_j \\ 0 & \text{si no} \end{cases}$$

Capítulo 3

Consumo de Marihuana

Varios han sido los intentos por establecer una teoría que explique el consumo substancias. Todas ellas, al igual que las variables que utilizan, tienen elementos en común, algunos elementos diferenciadores, y diferentes niveles de respaldo empírico. [68] es un intento por entender las similitudes, diferencias, intersecciones y vacíos de las distintas teorías más prominentes. Este enfoque será utilizado para dar orden a las variables relacionadas específicamente con el consumo de marihuana.

Este capítulo se separará en dos partes: estudio de predictores en el consumo de marihuana, y entorno social. En la primera se expondrán los resultados de una revisión bibliográfica de documentos que respaldan o desacreditan la relación de algunos atributos con respecto al consumo de marihuana. La segunda será destinada a detallar las medidas del entorno social que fueron utilizadas en los modelos predictivos. Estas resultan ser las variables más lógicas a utilizar en un contexto como el de *Twitter*, dado que las conexiones dentro de este medio son similares a las generadas por el contacto natural entre las personas.

3.1. Estudio de Predictores

La forma de presentar los atributos constará de tres grandes partes relacionadas con los tipos de influencia ejercidos sobre el individuo. Estos son: intrapersonal, cultural o referente a actitudes, y social o interpersonal. Además se expondrán otro tipo de características que están ligadas al uso de marihuana. La mayoría de los estudios han sido aplicados a adolescentes y adultos jóvenes, que han mostrado ser los más propensos a presentar el fenómeno estudiado. No se profundizará en los detalles metodológicos de cada documento.

3.1.1. Intrapersonal

Según [68], el tipo de influencia intrapersonal es conformado por aquellas características que están ligadas al individuo como tal. En esta categoría se encuentran rasgos de la personalidad,

disposiciones personales y estados de humor, habilidades de comportamiento, y creencias acerca de las habilidades propias con respecto a la substancia.

Motivacional

Los estudios que escrutaron variables motivacionales buscaban comprobar que el uso de substancias se origina desde un conjunto de motivaciones personales. En [77] se aplicó Análisis Factorial que apoyó el modelo motivacional de marihuana de 5 factores, los cuales resultaron ser predictores significativos para el uso de substancias. [98] también apoyó la hipótesis de los 5 factores, donde conjuntamente explicaban un 30 % de la varianza del uso de marihuana. En [58] exploraron otro tipo de motivaciones.

La importancia de los componentes motivacionales varía con respecto a la frecuencia de uso. En [10] se halló que sólo ciertos factores estaban únicamente relacionados con el uso actual y en [8] descubrieron que el factor de conformidad no explicaba significativamente el uso de los últimos 30 días. De igual manera, la importancia relativa de cada componente es diferente. En [10] apuntaban que los factores de mejora y social tienen relación positiva con el uso incrementado (9 % y 3 % de varianza única, respectivamente), y el factor conformidad tuvo relación inversa con dicho uso (5 % de varianza única). Además [76] señaló que el efecto de expansión es más importante en la marihuana que en el alcohol, argumentando razones diferentes para ingerir cada droga.

Algunas motivaciones están relacionadas con sentimientos. En [10] se observó que la ansiedad social era un predictor significativo para motivaciones de afrontamiento y conformidad, controlando por otras variables relevantes. Además el factor de afrontamiento mediaba la relación de ansiedad social y problemas con marihuana. [8] confirmó la importancia de la ansiedad social con afrontamiento, pero la negaba para todos lo demás factores. Además en el mismo estudio vincularon indirectamente a las expectativas de regulación de humor negativo con problemas de marihuana a través de motivos de afrontamiento.

Psicosocial

Las variables psicosociales residen en los procesos mentales de la persona e influyen en el comportamiento en respuesta al entorno (conducta). En efecto, tal como se indica en la sección de variables motivacionales y se menciona en [46], las variables de personalidad predicen riesgo de uso de sustancias. En algunos estudios pretendieron describir a miembros de un grupo en base a estas variables psicosociales y en otros a investigar el impacto de ellas sobre el uso de marihuana. La mayoría de veces su efecto está medido junto a variables sociales, pues se asocian intrínsecamente. Cuando se utilizan en conjunto pueden explicar hasta un 50 % de la varianza [46].

En algunas publicaciones ([42], [45]), se ha identificado que los consumidores comparten características comunes: son socialmente equilibrados, están abiertos a nuevas experiencias, y son impulsivos, buscan placer y son rebeldes. Además mayor involucramiento con la marihuana fue asociado con mayor apreciación por la independencia que por logros académicos, mayor tolerancia a la desviación (conductas fuera de lo socialmente aceptado), y apoyo a comportamientos problemáticos.

Evidencia encontrada en [41] apunta a que las variables psicosociales predicen jerárquicamente el consumo de marihuana, en otras palabras, actúan por medio de otras variables. Precisamente, sin tomar en cuenta el efecto de variables sociales, en [53] observaron que la personalidad desinhibida explicaba el consumo con ratios de probabilidad entre 2 y 3. En [87] hallaron que personas con experiencias de estrés postraumático y depresión eran más propensas a incrementar el uso de marihuana.

Las variables psicosociales pueden afectar directamente variables sociales. En [40], notaron que la búsqueda de sensaciones en menores predecía la interacción con pares desviados y niveles de imágenes sociales favorables de menores que usaban marihuana. A su vez, en [46] y [44] hallaron que el conjunto de patrones de personalidad y percepción del entorno social predecían el consumo actual de marihuana y su inicio dentro de los dos años posteriores.

3.1.2. Cultural o Referente a Actitudes

Según [68], son variables que enfatizan actitudes propias del individuo específicas a la sustancia y factores que afectan a aquellas actitudes. La mayoría de teorías que incorporan esta clasificación identifican a los valores personales como factores críticos en la formación de actitudes positivas hacia el uso de sustancias. Además se sostiene que la formación de dichos valores personales puede estar originada en el entorno que rodea al sujeto.

Evaluaciones Personales

En el marco diseñado en [68] se describe que estas variables se componen de un nivel de influencia intermedia como valores generales y comportamientos que contribuyen en las actitudes a las substancias, y un nivel de influencia próximo como creencias y evaluaciones sobre los costos y beneficios de cada droga. Las teorías más directas del consumo de substancias postulan que el fenómeno es explicado principalmente por estos condicionantes, considerando los demás como efectos indirectos.

Los valores generales han probado tener influencia significativa en el consumo de marihuana. En [95] se descubrió que para estudiantes universitarios, los objetivos de vida y académicos son predictores independientes del consumo en los últimos 90 días. Mientras más importantes son los objetivos menos probable es el consumo. [79] planteó que menos conflictos entre objetivos y el uso de marihuana estaba positivamente asociado al inicio y la frecuencia de uso, y en [74] se encontró una asociación indirecta positiva entre utilidad de uso y problemas relacionados con marihuana. Según [80], los objetivos identificados con mayor frecuencia fueron la regulación afectiva e interpersonal. Por otra parte, en [31] se halló conexión con la perspectiva del tiempo.

Las evaluaciones personales de la marihuana se estudiaron, por un lado en [97], desde la perspectiva de efectos subjetivos, donde ambas escalas (positivas y negativas) fueron positivamente conectadas con abuso o dependencia de marihuana. Por otro lado, en [81] se estudió el uso de marihuana desde la perspectiva de asociaciones en la memoria, arrojando predicciones significativas en el uso subsecuente.

Actitudes

En este apartado sólo se resumen los resultados de [75]. La actitud hacia una experiencia libre de drogas se identificó como moderador. Participantes con una actitud positiva hacia la marihuana y poca cercanía con la experiencia libre de drogas resultaron ser los individuos con más alta razón de consumo.

Percepción de Riesgo

La percepción de riesgo es una variable que históricamente ha sido reconocida como predictor del consumo de marihuana. Puede ser vista como el riesgo para las personas que significa consumir la droga. En estudios antiguos ([4], [3]) se trató de explicar las subidas o bajadas a nivel agregado en el consumo, usando la percepción de riesgo y factores de estilo de vida. En dichos estudios se reconoció que la percepción de riesgo podía explicar aquellas fluctuaciones.

La variable también ha sido escrutada para obtener relaciones más específicas. En efecto, en [51] se halló que la percepción de riesgo era mayor entre personas que no usaban marihuana frente aquellas que si lo reportaron. De manera más precisa, en [61] se concluye que la baja percepción de riesgo predice haber usado alguna vez, uso mensual, e intenciones futuras de usar marihuana. En [5] se explica que estudiantes quienes sintieron que los efectos físicos y psicológicos adversos no eran razones importantes para descontinuar el uso y aquellos que vivenciaron los efectos esperados en la etapa experimental eran los más propensos a continuar el consumo. Además se sostiene que cualquier relación entre factores sociales y el uso continuado es mediada por los efectos percibidos y el riesgo de la droga.

Otros estudios han realizado investigaciones para identificar variables que predigan la percepción de riesgo. Mayor edad y bajo nivel de conocimiento de los daños físicos y psicológicos de la droga fueron los predictores más fuertes encontrados en [61]. Finalmente, en [32] concluyeron que esta vez eran los factores sociales los que mediaban la relación entre la percepción de riesgo y el uso de marihuana.

3.1.3. Social o interpersonal

Según [68], las influencias interpersonales se enfocan en características y comportamientos de las personas que conforman el sistema de apoyo más íntimo del individuo. Además se centran en creencias del individuo acerca de lo común que es el uso de substancias, y las actitudes y comportamientos relacionados con la substancia de aquellas personas más cercanas.

Control y Relación de Padres

En varios estudios sostienen que las relaciones y el control por parte de los padres tienen influencia a lo largo de la vida su descendencia, aunque esta varía en función de la edad. En [25] se

mostró que la relación entre uso de pares y el uso adolescente de marihuana era atenuado por la cercanía con el padre y la percepción de ser atrapado por los padres. En [82] manifiesta algo similar: el apoyo por parte de los padres reduce el riesgo de uso de marihuana. Además que el monitoreo durante la adolescencia reduce el riesgo de consumo en la adultez temprana.

Redes Sociales

Estas variables sociales intentan incorporar el efecto de las conexiones dentro del entorno social. Según [68], considera características de las personas que constituyen el sistema de apoyo más íntimo del sujeto; el grado de apego emocional, y las actitudes y conductas de los modelos a seguir; y las creencias normativas acerca del uso y las presiones para consumir drogas. La Teoría de Aprendizaje Social fue una de las primeras en considerar los efectos sociales en el desarrollo de comportamientos y quizás la más conocida. Esta teoría define un proceso de adopción de conductas y tal como muestra [59], las variables se sostienen en un modelo estable de predicción, pero es un modelo limitado.

En [85] y [15] estudiaron factores de riesgo que incorporaban factores sociales y los consideraron como los predictores más fuertes en todas las etapas de consumo. En otro estudio donde comparaba varias teorías e indagaba el efecto conjunto de variables grupales, compromisos y psicológicas concluyeron que la orientación hacia un grupo de referencia usuario de marihuana es el predictor más sustancial de uso de marihuana ([36]).

En [91] se señaló que jóvenes con más usuarios de sustancias en sus redes sociales reportaron mayor consumo. Más precisamente, en [2], controlando por características de los padres y otros parámetros, encontraron que un incremento en el 10 % de amigos cercanos y compañeros de curso quienes usaban marihuana incrementaba la probabilidad de uso en un 5 %. [86] agrega que la influencia de los pares se sostiene en el periodo de crecimiento, pero la influencia de los padres disminuye con el paso del tiempo. Esto lo confirma [82], señalando que aquellos que se relacionaron con usuarios entre la adolescencia tardía y la adultez temprana eran 1.6 veces más propensos a iniciar el uso de marihuana. Además confirma la relación entre usuarios cercanos y uso propio para el inicio y continuación del fenómeno.

La posición dentro de la estructura de la red social también influye en la conducta. Efectivamente, [29] sostiene que adolescentes menos incrustados en la red, mayor estatus y mayor proximidad a pares usuarios de substancias eran más propensos al mismo comportamiento. En [52] se halló que el uso por parte del grupo, e interacciones entre la posición en la red y el uso de pares predicen el consumo. En particular, personas que conectan grupos son especialmente afectados por el consumo.

Normas Sociales

Las normas sociales son ligadas con la percepción del individuo de la aprobación de los pares sobre algún comportamiento. Aplicado a la marihuana, en [50] se han presenciado variaciones de desaprobación de la marihuana en distintos cortes generacionales y que estas diferencias afectan

al consumo. Los cortes con menos de la mitad de desaprobación evidenciaron probabilidades de consumo 3.53 veces mayor que en cortes con 90 % aprox. de desaprobación.

A nivel individual, [57] muestra que el nivel de aprobación personal es similar al nivel de aprobación de amigos cercanos, y que todos los grupos tienen una percepción similar de la aprobación del estudiante típico. Además un mayor uso de marihuana tiende a producir mayor aprobación personal, mayor aprobación percibida de los amigos cercanos y mayor aprobación por parte de los padres.

Aculturación

La aculturación pretende medir el nivel en que personas se han familiarizado en un ambiente inicialmente nuevo. Los estudios cubiertos sugieren relaciones diferentes en las etapas de uso. En [64] sostienen que la aculturación entre hispanoamericanos, medida a través del uso del lenguaje local, está asociada a riesgo incrementado de uso de marihuana alguna vez en la vida. Mientras que en [69] indican que entre el mismo grupo de estudio la aculturación fue asociada a bajo uso de marihuana. Además se incluye que el monitoreo de los padres puede mediar este efecto.

3.1.4. Otras variables

En esta categoría se aglomeran todas aquellas variables que mencionan otro tipo de efectos en el uso de substancias. Eventos en el transcurso de la vida, cortes generacionales, edad, uso de otras drogas, nivel socioeconómico, entre otras, son características que afectan de alguna manera al grupo de personas que rodea al sujeto de estudio y al mismo. Muchas veces son ocupadas como variables de control para probar la importancia del grupo de variables mencionados a lo largo del capítulo.

Edad y Eventos

En [47], [55], [90] y [48] se han determinado las etapas de la vida más propensas al consumo de marihuana. Se ha detectado que el riesgo de iniciación cubre toda el desarrollo del adolescente, porque están expuestos a otros quienes usan marihuana. El riesgo de iniciación en la población de estudio comenzaba a subir a los 13 años, alcanzando su punto máximo a los 18 años, después de esta edad se estabiliza. El punto de mayor uso también se encuentra en los 18 años. Por otro lado, se ha detectado que mientras más tarde se inicia el consumo menos probable es incrementar su frecuencia.

En [28], [71] y [83] se identificaron trayectorias diferentes para grupos que compartían características en común. El número de grupos identificados es diferente para cada estudio, pero todos estaban de acuerdo que existen distintos tipo de usuarios. Los más generales son abstemios, ocasionales, crónicos y otros que tienen diferentes pendientes en la frecuencia de uso a lo largo de su vida.

Los estudios [92] y [82] probaron la relación de asistir a la universidad. En general la relación existe para ambos lados, relación positiva y negativa, dependiendo de otros factores. Por otra parte, [96], [96] y [82] hallaron que el embarazo, el matrimonio y la convivencia tienden a disminuir las probabilidades de uso de cannabis.

Factores Socioeconómicos

La relación de los factores socioeconómicos y el consumo de marihuana es menos claro, ya que existe evidencia que soporta la influencia positiva y negativa ([82]). En otro estudio ([63]) concluyeron que los efectos de los factores socioeconómicos no eran significativos si se controlaba por otras variables como el uso por parte de amigos.

Puerta de Entrada o "Gateway"

El consumo de diferentes drogas se ha relacionado históricamente, argumentando la necesidad de utilizar drogas más fuertes para obtener efectos más pronunciados y que el ambiente de cada droga genera oportunidades para consumir otros tipos de sustancias. Los estudios [82], [37], [26], [89] y [93] explican que estas relaciones existen efectivamente. Aunque se encuentran diferencias con respecto a los cortes generacionales. Se han asociado progresiones desde tabaco y alcohol hacia la marihuana, pero se sostiene que la conexión del tabaco es más marcada.

Genética

Esta variable ha sido evaluada con respecto a efectos directos e indirectos en el consumo de marihuana, siendo más comunes los segundos. En [62] se ha encontrado pruebas significativas que los factores genéticos influyen en algunos rasgos de la personalidad (por ejemplo, actitud de toma de riesgos) que a su vez afecta al consumo de marihuana. En [33] se halló que factores hereditarios incorporaban del 50 % de la varianza del consumo de marihuana en la vida, pero luego de controlar por los factores hereditarios de Trastornos de Personalidad Antisocial (TPA) en la vida bajó notablemente la varianza explicada. Además luego de este control, los efectos de la depresión sobre el uso de marihuana no fueron significativos.

Legalización

Esta variable es relativamente nueva, ya que la legalización es una medida que se ha tomado por sólo algunos estados en el último tiempo, haciendo difícil la evaluación de este factor. El estudio realizado a 50 estados en [11] indicó que los residentes de estados con leyes de marihuana medicinal tenían probabilidades mayores de evidenciar su consumo y dependencia que residentes de estados sin esas leyes. Estados que legalizaron la marihuana medicinal también tenían mayores tasas de uso. Aunque se desconoce la causalidad de las variables.

Género, raza, participación religiosa, educación y características de los padres

En una revisión de factores de riesgo y protectores ([82]) se resumió que en general los hombres están más propensos a la iniciación, consumo frecuente y problemas relacionados con la marihuana. Precisamente, los hombres fueron asociados 1.4 veces más que las mujeres al consumo. En referencia a la raza, se encontró que blancos no hispánicos, afro-americanos e hispánicos eran más propensos a desarrollar desórdenes de consumo. En la adolescencia se halló que era más probable que europeo-americanos desarrollaran problemas de uso en lugar de afro-americanos. El involucramiento religioso tiene evidencia de representar un factor de protección con respecto al uso de marihuana, al igual que mejores calificaciones en la edad escolar. Finalmente, mayor educación de la madre y cambios en su estatus marital fueron asociados a mayor probabilidad de consumo por parte de los hijos.

3.2. Entorno Social

Una vez hecha la revisión de los atributos, se determinó que las variables sociales tienen mayor factibilidad de ser replicadas para el contexto de *Twitter*, debido a la similitud que comparten con las conexiones tradicionales entre personas. Además el modelo más completo es el utilizado en [29], donde emplean medidas de Análisis de Redes Sociales. El modelo incorpora tres grupos de variables: incrustación social, estatus social y proximidad social a consumidores. Los resultados reflejaron que el consumo de marihuana es afectado por los tres grupos de variables. A continuación se detalla cada grupo con sus componentes:

3.2.1. Incrustación Social

El objetivo de las medidas de Incrustación Social es evaluar el posicionamiento y las características estructurales de un nodo dentro de su red social. Específicamente, evalúan la composición de su red egocéntrica, es decir, los nodos que están conectados directamente con él. La medidas evaluadas fueron las siguientes:

- Reciprocidad de nominaciones: esta medida contrasta aquellas nominaciones que fueron correspondidas de aquellas que no.
- Densidad de vecindario: es definida como el número de conexiones existentes dividida por el total de conexiones posibles.
- Posición social: mide la posición en la red dividiendo los nodos en tres grupos: miembros, puentes y aislados. Los miembros se caracterizas en que una remoción de algún nodo no divide al grupo. Los puentes tiene conexiones a grupos diferentes pero no son miembros de ninguno. Los aislados tiene una o ninguna conexión.
- Nominaciones externas: considera el número de nominaciones fuera de la red social. Mientras más nominaciones externas tenga un nodo, menos incrustado va a estar en la red social.

3.2.2. Estatus Social

Las también llamadas medidas de centralidad buscan reflejar la importancia de un nodo basada en la posición dentro de la red social. Muchas de las medidas evalúan la importancia con respecto a todos los nodos. Las medidas consideradas son:

- *Normed indegree*: es un indicador de popularidad definido por el número de nominaciones dividido por el total de nominaciones posibles.
- *Reach centrality*: mide el porcentaje de la red social que puede ser alcanzado mediante tres o menos nominaciones.
- *Betweeness centrality*: este indicador es construido calculando primero los caminos geodésicos entre todos los nodos de la red. Luego, es calculado el número de caminos donde el nodo pertenece. Refleja el poder del nodo como regulador de flujos de información.
- *Bonacich's power centrality*: para un nodo, esta medida de centralidad es calculada ponderando la centralidad de los nodos adyacentes. Un nodo tendrá mayor Bonacich's power centrality si sus amigos tienen mayor centralidad.

3.2.3. Proximidad Social a Consumidores

Estas medidas simplemente evalúan la cercanía de un nodo a otros con el comportamiento. Son usadas diferentes versiones, pero la idea detrás es la misma. Las medidas usada son:

- Consumo del mejor amigo: evalúa la posibilidad de que el mejor amigo reporte consumo.
- Número de consumidores en el vecindario: realiza sumatoria de todos los nodos adyacentes que reporten consumo, sin considerar al nodo mismo.
- Distancia al consumidor más cercano: el nombre es sugerente, mide los saltos entre nodos para llegar al primer consumidor.

Capítulo 4

Diseño

En este capítulo se trazarán las líneas para implementar el sistema que permita reconocer el consumo y opinión de marihuana por parte de usuarios chilenos de *Twitter*. En primer lugar, se expondrán los requerimientos y expectativas del cliente. Luego, se expondrá el tipo de información que fue usado y su estructura, así como los detalles correspondientes a la disposición de datos para el entrenamiento de los algoritmos. Luego, se precisarán las formas y medidas para evaluar el rendimiento de los distintos clasificadores. Finalmente, se enumerarán los módulos esenciales para el funcionamiento de la aplicación y operaciones principales.

4.1. Requerimientos

En esta sección se diseñarán las métricas para caracterizar el consumo y la percepción de la marihuana en *Twitter*. El supuesto básico que apoya la proposición de estas métricas consiste en que el Estudio Nacional de Drogas (realizado por el SENDA) incorpora aquellas variables necesarias para caracterizar el consumo de drogas en la población, en particular para la marihuana. Cabe destacar que aquellas métricas son utilizadas como un punto de referencia para diseñar nuevos indicadores, ya que la naturaleza de la información de *Twitter* hace imposible replicar los mismos datos que son recolectados por el estudio original.

A continuación son listadas las principales variables incorporadas por este estudio, que a su vez, son segmentadas por dos tipos de variables (socio-demográficas y geográficas).

4.1.1. Variables Originales

Las principales variables incorporadas por el Estudio Nacional de Drogas para Población General, cuyo objetivo es describir la magnitud y tendencias en el consumo de drogas lícitas e ilícitas son mostradas en las siguiente lista:

1. Prevalencia (Porcentaje de población consumidora).

2. Percepción de riesgo.
3. Percepción de facilidad de acceso.
4. Ofrecimiento reciente de drogas.
5. Lugar de último ofrecimiento de marihuana.
6. Percepción de drogas en el entorno de la casa.
7. Tipos de marihuana conseguidas con mayor frecuencia.
8. Dependencia entre prevalentes.
9. Consumo de riesgo de alcohol entre prevalentes de alcohol.
10. Uso intenso de alcohol y máximo número de tragos.
11. Intensidad de uso de drogas.
12. Percepción de problemas de barrio de residencia.
13. Aprobación de políticas de control de alcohol.
14. Aprobación de políticas de control de drogas.
15. Consumo de sustancias y conducción.

4.1.2. Segmentación Original

La segmentación es realizada en base a variables de dos tipos: socio-demográficas y geográficas. En sumatoria son utilizadas cuatro variables de segmentación, cada cual es medianamente auto-explicativa.

1. Sexo: la categoría se divide en hombres y mujeres.
2. Edad: las personas se clasifican según cuatro rangos etarios (12-18, 19-25, 26-34, 35-44 y 45-64).
3. Nivel Socio-económico: se dividen en personas de nivel socio-económico bajo, medio y alto, según el nivel de ingresos declarado.
4. Regiones: se subdivide por las 15 regiones administrativas de Chile.

4.1.3. Selección de métricas y segmentación

La lista potencial de variables fue seleccionada y priorizada en base a tres criterios. Dichos criterios son resultado de restricciones y necesidades del cliente y el desarrollador de esta aplicación:

- Factibilidad: posibilidad técnica.
- Importancia: interés desde el punto de vista del cliente.
- Alcance: plazos del proyecto.

Fueron extraídas las siguientes variables, el orden corresponde a la importancia relativa asignada por el cliente a cada una de ellas:

1. Prevalencia: porcentaje de la muestra que reconoce haber consumido marihuana dentro del periodo evaluado. La encuesta abarca tres ventanas de tiempo: durante la vida, último año y último mes.

2. Percepción de riesgo: corresponde al porcentaje de la muestra que considera que consumir marihuana representa un riesgo grande. Para la marihuana, el ítem estima consumo frecuente y experimental (una o dos veces).

3. Percepción de facilidad de acceso: porcentaje de la muestra que considera fácil conseguir marihuana.

4. Ofrecimiento reciente de drogas: proporción de la muestra que declara haber recibido oferta de marihuana durante los últimos doce meses.

5. Tipos de marihuana: tipo de marihuana consumida con mayor frecuencia. Considera como alternativas posibles la marihuana prensada o "paraguaya", marihuana verde o yerba, y marihuana “skunk” o “transgénica”.

6. Aprobación de políticas de control de drogas: porcentaje de individuos que se declara muy de acuerdo o de acuerdo con respecto a diferentes medidas en el control de drogas, tales como permitir el uso de marihuana con fines terapéuticos, entre otras.

7. Intensidad de uso: promedio de días de consumo en el último mes.

Por otra parte, se identificó como requerimiento del cliente sólo segmentar por rango etario, ya que corresponde al dato más importante a nivel de intervención. La dificultad de replicar el total de las variables de segmentación en el contexto de las redes sociales amerita esta priorización, dado que estas no están explícitamente declaradas.

4.1.4. Indicadores Finales

Debido a la naturaleza de la información presente en *Twitter*, fue necesario realizar adaptaciones a las variables antes mencionadas. Cada transformación puede resultar en un indicador que difiere totalmente de la naturaleza de su variable inspiradora, pero cumple con el propósito principal de

brindar información que ayude a comprender el comportamiento de las personas con respecto a la marihuana. La totalidad de indicadores fueron validados por el cliente y algunos tienen el objetivo de ampliar la frontera de posibilidades actuales.

1. Polaridad de *Tweets*: está diseñado para obtener información acerca de la opinión sobre la marihuana. Es un indicador que brinda conocimiento de qué tan positivo o negativo es un *tweet*, lo cual podría estar relacionado con la percepción de riesgo de la droga. Se calculará con una aplicación de *Opinion Mining* y se obtendrán medidas agregadas.

2. Prevalencia: tal como su variable inspiradora, este indicador busca medir el porcentaje de individuos que consume marihuana. Será determinado directa o indirectamente por la información disponible en su cuenta, sus Tweets publicados y su contexto social.

3. Porcentaje de "amigos" usuarios: es una medida que se espera que pueda ofrecer nociones de facilidad de acceso de cada individuo. Ya que se ha demostrado que uno de los focos principales de acceso a las drogas son personas pertenecientes al circulo cercano [68].

4. Oferta de marihuana: existen estudios que han confirmado la venta de drogas vía Internet [18]. Por esto, se espera poder hallar en Twitter cuentas que ofrecen sus productos públicamente y de esta forma, se pretende obtener un indicador que asemeje la variable de oferta reciente de marihuana.

5. Palabras utilizadas: se espera realizar una categorización de las palabras utilizadas para describir el consumo de marihuana como una forma de determinar el tipo de marihuana.

6. Polaridad de Tweets de políticas: al igual que el primer indicador, este busca recoger los sentimientos de los comentarios acerca de marihuana, pero para aquellos Tweets referidos como políticas de control.

7. Frecuencia de uso: este indicador está diseñado para incorporar información relacionada con la frecuencia de Tweets de los individuos que implican directa o indirectamente el consumo de marihuana.

En lo que sigue del capítulo se delineará el tipo de información y la forma de brindarle estructura, la evaluación de rendimiento de las técnicas de Data Mining y los componentes de la aplicación que hacen posible darle forma a los indicadores que fueron descritos en esta sección.

4.2. Descripción de Datos

En esta sección se describirá el tipo de información que será utilizada como datos de entrada para la aplicación. En primer lugar se profundizará en la información extraíble por medio de la API de Twitter, moldeada por la naturaleza del servicio. Luego, se establecerá la estructura de datos requerida para el entrenamiento de los algoritmos de aprendizaje. Finalmente, se definirá por separado los mecanismos de etiquetado de textos y de usuarios para sus respectivos análisis y algoritmos de clasificación.

4.2.1. Datos disponibles

En una fase inicial del proyecto se consideró la utilización de varias fuentes para recolectar datos. En la práctica, fueron evaluados tres sitios de intercambio de información entre usuarios: *Twitter*, *Facebook* y un sitio llamado AmigosdelCannabis. Esta evaluación resultó en la elección de una sola fuente: *Twitter*. La razón de esta acción se sostiene en tres criterios:

- Acceso: la información de los usuarios es un asunto delicado para la mayoría de las redes sociales. Por esta razón implementan políticas de privacidad que restringen el acceso a los datos. Este es el caso de *Facebook*, donde se puede obtener información limitada aún siendo miembro de la comunidad o utilizando su API. Utilizando la API es posible obtener información pública del usuario, pero es imposible adquirir las conexiones entre usuarios, datos esenciales para los propósitos de este proyecto.
- Utilidad: este criterio se aplicó en el caso de AmigosdelCannabis. Similarmente al criterio anterior, es casi imposible obtener información de las conexiones entre usuarios desde este sitio. La razón no se sostiene en el acceso restringido, sino en la inexistencia. Esto descalifica al sitio como fuente útil de información.
- Disponibilidad: a diferencia de los criterios anteriores, este criterio favorece la elección de una alternativa. La preferencia por *Twitter* se fundamenta en la disponibilidad de otra fuente de datos previamente construida y constantemente actualizada por el equipo de *Opinion-Zoom*. Esta base de datos favorecerá la prueba y desarrollo de algoritmos y procesos.

La información disponible en *Twitter* está moldeada por las funcionalidades que ofrece su servicio de microblogging. Para efectos de este proyecto, la información útil se puede dividir en tres tipos:

- Información acerca del usuario.
- Red del usuario formada por sus conexiones con otros usuarios.
- *Tweets* publicados por el usuario.

La API REST de *Twitter*, permite el acceso a cada tipo de información en la lista. Para acceder a ella, sólo es necesario conocer el identificador de cada usuario (número de identificación o nombre de usuario) o si se desea buscar un *Tweet* en específico, el identificador de este (número de identificación). El conjunto de datos es entregado en formato JSON. En la Tabla 4.1, se puede observar la estructura de los datos del usuario. Aunque en su totalidad, la cantidad de datos es mayor, aquí son expuestos sólo los datos útiles para la construcción de los indicadores.

Los *tweets* de un usuario serán obtenidos de dos formas. La primera es en forma de lista, es decir, obtener un conjunto de *tweets* asociados a un usuario y la segunda, el *tweet* individual asociado a una ID. Los datos asociados a un *tweet* son mostrados en la Tabla 4.2. Al igual que en el caso del usuario, son considerados sólo los datos utilizados en este proyecto.

La información de la red del usuario se entrega en forma de listas. Una lista tiene la ID de otros usuarios como único tipo de dato. Estas listas son de dos tipos. Una entrega el registro de todos los usuarios a quien sigue (amigos) y la otra entre el registro de usuarios que le siguen (seguidores).

Nombre	Descripción	Tipo
ID	número identificador	entero largo
screen name	nombre identificador	texto
name	nombre del usuario	texto
description	auto-descripción	texto
lang	lenguaje	texto
location	ubicación	texto
timezone	zona horaria	texto
createdat	fecha y hora de creación de la cuenta	tiempo
followerscount	número de seguidores	entero
friendscount	número de amigos	entero
statusescount	número de tweets	entero
isgeoenabled	opción de geo-localización	boolean
isprotected	tweets protegidos	boolean

Tabla 4.1: Datos del usuario

Nombre	Descripción	Tipo
ID	número identificador	entero largo
user	usuario asociado	entero largo
text	contenido del tweet	texto
istruncated	mensaje cortado	boolean
retweetcount	número de veces republicado	entero
createdat	fecha y hora de creación	tiempo

Tabla 4.2: Datos del *tweets*

4.2.2. Estructura Necesaria

La estructura necesaria de los datos está dada por la naturaleza de los indicadores y la forma de construirlos. Los tratamientos y metodologías que se necesiten aplicar para llegar a ella están dados por los datos que se dispongan de entrada. Tal como se vio en la sección anterior, se puede deducir que se cuenta con datos estructurados y no estructurados. En primer lugar cada usuario tiene la mayoría de sus características bien estructuradas, pero hay varios campos que contienen textos. En este marco, los textos por definición no poseen estructura y deben ser tratados especiales para aplicar técnicas de reconocimiento de patrones. Esto mismo pasa con los *tweets*, siendo su campo más importante el *tweet* en sí mismo.

El enfoque de este proyecto hace énfasis de la necesidad de transformar los tweets en información relevante y que pueda ayudar a la toma de decisiones. Para hacer esto, en primer lugar se requiere clasificar a los tweets con respecto a tres acciones:

- Consumo de marihuana (binaria).
- Mención de política de control de marihuana (binaria).
- Venta de marihuana (binaria).

En segundo lugar, se pretende utilizar los *tweets*, la información del usuario y análisis de redes sociales para determinar dos características en el usuario:

- Consumo de marihuana (binaria).
- Rango etario (categórica).

Para llevar a cabo las clasificaciones antes mencionadas se utilizarán algoritmos de aprendizaje supervisado, donde es necesario tener un conjunto de casos previamente etiquetados (manualmente). Además en el caso del texto, se requiere la construcción de un conjunto de variables que representen a cada documento. Las Tablas 4.3 y 4.4 muestran ejemplos de los etiquetados necesarios para *tweets* y para usuarios, respectivamente. Los casos de etiquetado de tweets y usuarios serán vistos en profundidad en las dos secciones que siguen. En la sección de diseño de la aplicación se revisará en detalle el preprocesamiento de textos (columnas de variables en la Tabla 4.3).

	Variables					Etiquetas		
Tweet	1	...	k	...	m	Consumo	Política	Venta
1	1	...	1	...	0	1	0	0
2	0	...	1	...	0	1	1	0
⋮	⋮	⋱	⋮	⋱	⋮	⋮	⋮	⋮
j	0	...	0	...	1	0	0	1
⋮	⋮	⋱	⋮	⋱	⋮	⋮	⋮	⋮
n	1	...	1	...	0	0	0	1

Tabla 4.3: Etiquetado de *tweets*

	Variables					Etiquetas	
Usuario	1	...	k	...	m	Consumo	Edad
1	1.3	...	11	...	0	1	18
2	0.2	...	20	...	0	1	35
⋮	⋮	⋱	⋮	⋱	⋮	⋮	⋮
i	2.2	...	5	...	1	0	22
⋮	⋮	⋱	⋮	⋱	⋮	⋮	⋮
n	-0.1	...	4	...	0	0	15

Tabla 4.4: Etiquetado de usuarios

4.2.3. Etiquetado de *Tweets*

La extracción de información desde textos no es una tarea difícil para el humano, pero la misión se dificulta cuando el número de documentos aumenta drásticamente. Por esta razón, se hace necesario el uso de algoritmos de entrenamiento que reconozcan parámetros en el texto e idealmente clasifiquen con el mismo grado de precisión o mejor que la mente humana. En el caso de los algoritmos de aprendizaje supervisado es imperioso tener un conjunto de entrenamiento, el cual debe ser etiquetado manualmente por personas.

Con el fin de obtener un etiquetado consistente de los documentos, se diseñó un conjunto de reglas de etiquetado de textos, las cuales son enumeradas a continuación:

1. El evaluador debe etiquetar el documento en respuesta a una pregunta definida claramente para cada categoría.

2. Cada una de las categorías del documento será evaluada en la misma instancia (por *tweet*).

3. Cada documento debe ser etiquetado por sólo una persona que califique como experto (usuario de *Twitter*).

4. Se seleccionará un porcentaje de casos que será etiquetado por todas las personas. Para ese conjunto se determinará el índice Kappa de Cohen, el cual indica la concordancia entre etiquetadores.

Fue seleccionado un grupo de 1.500 *tweets* para el proceso de etiquetamiento. Este número tiene asociado un error del 3 % y un 98 % de confianza. Este conjunto de *tweets* será clasificado por un grupo de 12 personas. Cada uno de los 1.450 textos será etiquetado por sólo una persona y 50 por el grupo entero. Resultando en alrededor de 171 *tweets* por persona. Además la división y la distribución de textos será hecha al azar.

4.2.4. Etiquetado de Usuarios

A diferencia del etiquetado de *tweets*, el etiquetado de usuarios no se ejecutará por parte de terceros. Si bien es un mecanismo confiable que permite el etiquetado de gran cantidad de docu-

mentos, se lleva a cabo con asimetrías de información. En esta instancia se busca la veracidad de las etiquetas. Por esta razón, dichas etiquetas serán asignadas por los mismos usuarios bajo evaluación. La necesidad de esta metodología diferente se determina por el significado del indicador final. Si se usara sólo el etiquetado de terceros, se obtendría un indicador de aquellos que mencionan abiertamente su consumo. Pero se busca un indicador de prevalencia "real", calculado en base a variables originadas en los *tweets* y por la red social de *Twitter*. Nuevamente, el indicador estará acotado al contexto del servicio de *microblogging*.

El etiquetado se desarrollará mediante una encuesta directa a los usuarios. Una vez sea construida una base de datos de usuarios chilenos, se escogerán casos al azar para enviarles la encuesta. La cual será publicada en un *tweet* para cada usuario, mencionándole directamente. La encuesta debe contener preguntas que permitan determinar el consumo de marihuana, la edad y el sexo de cada usuario. Además el usuario debe ingresar el nombre de usuario de *Twitter* para relacionar los resultados de la encuesta con la base de datos.

4.3. Evaluación de Rendimiento

Esta sección abordará la evaluación de rendimiento de los clasificadores usados para el funcionamiento de la aplicación. Se debe aceptar que los indicadores resultantes no serán totalmente confiables, porque los clasificadores que los sostienen no lo son. Se debe asumir cierto grado de error y esta sección precisará la forma de traducirlo en números.

Las clasificaciones consideradas por este proyecto forman parte de los casos típicos, por ende la evaluación de rendimiento está ampliamente estudiada. A pesar de esto, cada problema amerita evaluar qué algoritmo es el que produce mejores resultados. Para hacer esto, se debe determinar las métricas que se emplearán para realizar la evaluación de rendimiento. Por otro lado, la elección de las métricas está condicionada por los resultados esperados de los algoritmos. En el caso de la aplicación, se prefiere por parte del cliente, obtener pocas predicciones confiables que gran cantidad de predicciones poco confiables.

Generalmente, la métricas de evaluación de rendimiento en problemas de clasificación son derivadas desde una matriz con los número de casos clasificados correcta e incorrectamente para cada clase, llamada matriz de confusión. La Tabla 4.5 muestra un ejemplo de esta matriz para el caso de dos clases. Es posible generalizar para mayor número de clases. Los valores dentro de la matriz se describen a continuación:

- *Falso Positivo (FP)*: casos predichos como positivos, que verdaderamente tienen un valor negativo.
- *Falso Negativo (FN)*: casos predichos como negativos, pero que tienen valor positivo.
- *Verdadero Positivo (VP)*: casos predichos correctamente con valor positivo.
- *Verdadero Negativo (VN)*: casos predichos correctamente con valor negativo.

Las métricas de evaluación de rendimiento serán tres:

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

Tabla 4.5: Matriz de Confusión

- *Precision*: estima la probabilidad de que una predicción positiva sea correcta.

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

- *Recall*: es la proporción de casos que tienen un valor positivo y fueron correctamente predichos como positivos.

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

- *F-measure*: es una combinación entre Precision y Recall, donde una constante β controla el trade-off entre ambas métricas.

$$F - measure = \frac{(\beta^2 + 1) * P * R}{\beta^2) * P + R}$$

Aunque se calcularán las tres métricas de rendimiento, se priorizará el resultado de la primera, por la razón antes expuesta.

4.4. Diseño de Aplicación

En esta sección se detallará el diseño de la aplicación que hará frente a los requerimientos previamente establecidos, en otras palabras, la aplicación que hará posible la construcción de los indicadores. El diseño de la aplicación se confeccionará bajo el paradigma de programación modular. El cual separa las funcionalidades de la aplicación en componentes desarrollados independientemente, llamados módulos.

Se recomienda utilizar programación modular para toda aplicación, sin importar el nivel de complejidad, ya que hace uso de uno de los principios básicos para resolver problemas: "dividir para reinar". El problema de programación se subdivide en sub-problemas que pueden ser resueltos de manera sencilla. A su vez, cada módulo puede ser subdividido en tareas aún más simples. Bajo este enfoque, los módulos son ensamblados en una estructura que establece relaciones de dependencia claras. Según Boudreau en [9], la modularidad brinda diseños más claros y control de interdependencias de módulos, provee a los desarrolladores de más flexibilidad en el mantenimiento, y más beneficios se vuelven claros durante el crecimiento de la aplicación.

En el contexto del proyecto, la programación modular es necesaria para separar claramente las funcionalidades de la aplicación, las cuales son de naturaleza diferente y son utilizadas en ventanas

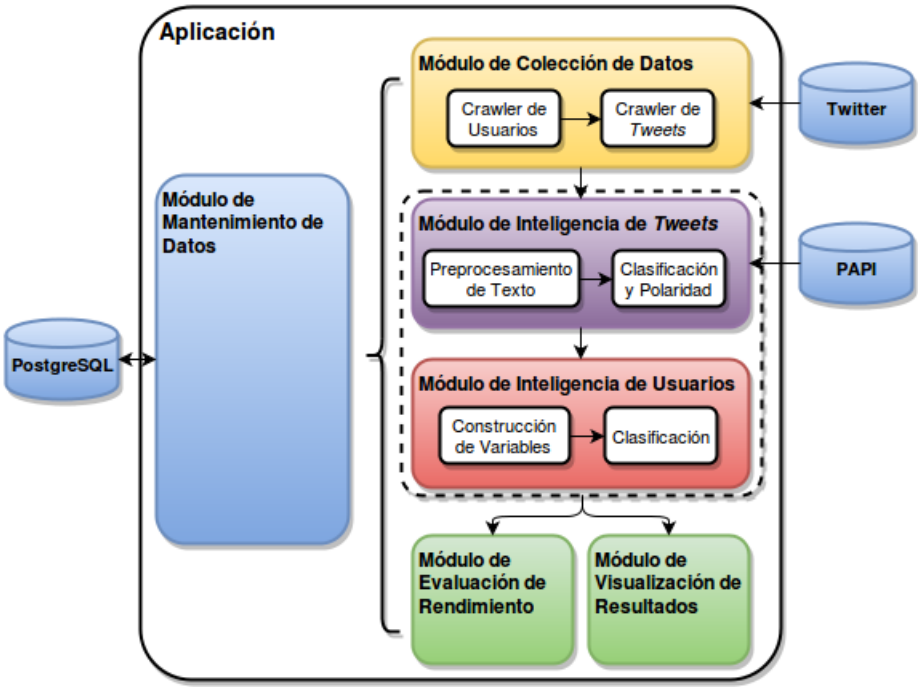

Figura 4.1: Diagrama Modular de la Aplicación
Fuente: Elaboración Propia

de tiempo distintas. Además se consideran futuras intervenciones en el código, por ende, se debe asegurar el entendimiento de la aplicación por terceros.

La estructura general de la aplicación será conformada por seis módulos: Modulo de Recolección de Datos (MCD), Módulo de Mantenimiento de Datos (MMD), Módulo de Inteligencia de *Tweets* (MIT), Módulo de Inteligencia de Usuarios (MIU), Módulo de Visualización de Resultados (MVR) y Módulo de Evaluación de Rendimiento (MER). El diagrama general del diseño se puede ver en la Figura 4.1. Todos estos módulos serán descritos con detalle en lo que queda del capítulo.

4.4.1. Recolección de Datos

El Módulo de Recolección de Datos (MCD) cumplirá la tarea de extraer la información que permite el funcionamiento de toda la aplicación, transformándola en una tarea primordial para los objetivos del proyecto. Para hacer esto se utilizará la API REST de *Twitter*, precediendo primero a extraer conjuntos de usuarios y en una segunda fase, extraer todos los tweets asociados a aquellos usuarios. Es necesario enfatizar esta separación, ya que la actualización de cada tipo de dato se hará en frecuencias diferentes. Luego se explicarán con mayor profundidad los fundamentos para obtener información de esta manera. En la Figura 4.2 se resumen estos dos casos de uso del módulo.

La API REST de *Twitter* asigna restricciones a las consultas, las cuales son procesadas sólo si tienen como origen a un aplicación válida. La validación se hace mediante un conjunto de cuatro series de caracteres que contienen la información del usuario asociado a la aplicación, la aplicación y claves codificadas. Todo esto se obtiene desde el sitio destinado a administrar aplicaciones de Twitter. Las restricciones limitan el número de peticiones por una ventana de 15 minutos para

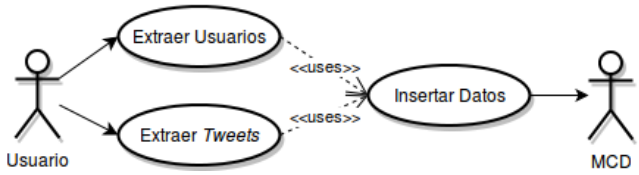

Figura 4.2: Casos de uso de MCD
Fuente: Elaboración Propia

Consulta	Límite
Información del usuario	180
Lista de seguidores por usuario	15
Lista de amigos por usuario	15
Lista de tweets por usuario	180
Estado de credenciales	180

Tabla 4.6: Restricciones de la API de Twitter

cada credencial. Los límites para cada consulta son mostrados en la Tabla 4.6. En él se puede apreciar diferencia entre las cifras, agregando complejidad al momento de diseñar el MCD. Para superar el inconveniente se utilizará una cola iterativa, el cual permitirá emplear ordenadamente las credenciales habilitadas en la ventana de tiempo.

La Figura 4.3 muestra el funcionamiento de la cola iterativa. En primer lugar se obtiene el estado actual del primer elemento de la cola. Si está habilitado para realizar consultas es ingresado al servidor, en caso contrario se espera el tiempo remanente de la ventana de tiempo de 15 minutos. Una vez que es utilizado el elemento (credenciales) es puesto en la parte posterior de la cola, dándole el nombre a la cola.

El desarrollo de algoritmos de recolección de datos se apoyará en sus inicios en una base de datos ya establecida del proyecto OpinionZoom, llamada *La Gorda*. Esta base de datos cuenta con gran cantidad de *tweets* asociados a un set de usuarios, en su mayoría chilenos. Los *tweets* de *La Gorda* han sido recolectados desde el 10 de Abril del año 2015 hasta la actualidad mediante la API Streaming de *Twitter*. La cantidad de usuarios seguidos por La Gorda (1,013,005 usuarios) y el número de *tweets* almacenados diariamente (1,707,496 aproximadamente) la transforman en una base de datos ideal para realizar pruebas y utilizar datos consignados en ella.

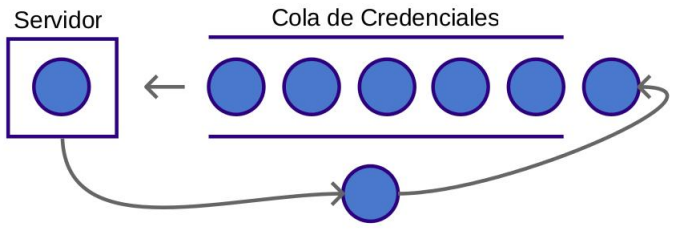

Figura 4.3: Cola de Credenciales
Fuente: Elaboración Propia

El enfoque de Twitter es la difusión de mensajes cortos a través de la red de usuarios, conectados a través de las relaciones de seguimiento. Las cuales no son más que enlaces direccionados de un usuario (nodo) a otro. Los objetivos de este módulo son obtener la red de usuarios chilenos, es decir, el conjunto de usuarios chilenos y sus conexiones, y los tweets publicados por cada uno de ellos. Las dos tareas se desarrollarán con frecuencias distintas, siendo mayor la segunda, debido a tres razones:

- Los recursos computacionales son limitados.
- Los indicadores de los *tweets* necesitan seguimiento continuo.
- La cantidad de *tweets* es más dinámica que la de usuarios.

El algoritmo de obtención de usuarios operará como uno de los algoritmos clásicos en recorrido de grafos, denominado Búsqueda en Anchura. La noción detrás del algoritmo es la siguiente: para cada elemento en la red se agregan todos los elementos adyacentes a él. El proceso se implementará haciendo uso de una lista y una cola. La lista permite dar seguimiento de los nodos ya agregados a la cola y la cola mantiene el orden de los nodos por visitar. El procedimiento inicia con una lista de elementos, llamada semilla, la cual será ingresada a la cola. Para cada elemento extraído de la cola, será agregado a la cola y a la lista todo aquel elemento que posea un enlace direccional desde él al nodo evaluado y que no haya sido agregado antes. Al final de la procedimiento se deberían haber visitado todos los elementos del grafo, a menos que el grafo general no sea conexo y la semilla no posea nodos de todos los elementos conexos.

El procedimiento descrito arriba será adaptado ligeramente para asemejar a un *Web Crawler*, llamado Crawler Focalizado. El cual recorre el grafo de la misma manera, pero sólo son agregados los nodos adyacentes de aquellos elementos que cumplan con cierto criterio. En el caso de la aplicación, el criterio consiste en que los usuarios sean chilenos, cuya información está contenida en los datos del usuario. Para operar de esta forma, es imperioso realizar consultas para obtener los datos de usuario y la lista de seguidores. La elección de esta lista es tomada porque los usuarios tienden a seguir a cualquier persona en cualquier círculo de relaciones que sea de su interés, pero son seguidas por un grupo más acotado, generalmente compuesto por su círculo cercano de conocidos.

En la Figura 4.4 se muestra el proceso iterativo de extracción de usuarios y en la Figura 4.5 se muestra un ejemplo de dos iteraciones. El algoritmo comienza con una semilla en la Figura 4.5.a, en las figuras 4.5.b y 4.5.d se incorporan los nodos adyacentes que cumplen el criterio (nodos verdes) y en las figuras 4.5.c y 4.5.e se muestra el estado final de cada iteración.

Se puede deducir que la aplicación del criterio de nacionalidad (ubicación estable) es crucial para el resultado final del "crawler", por lo tanto debe ser definido detalladamente. El campo *location* contiene información acerca de la ubicación del usuario y es usado para determinar su "nacionalidad". Aunque el campo es rellenado manualmente por el usuario, teniendo completa libertad a la hora de escribir (por ejemplo, "la luna"), en la práctica es el método más sencillo para determinar la ubicación estable del usuario. Se buscará que el campo contenga palabras correspondiente a las siguientes opciones:

- Nivel nacional: Por defecto, la palabra "Chile".
- Nivel regional: Los nombres de las 15 regiones administrativas de Chile (por ejemplo, me-

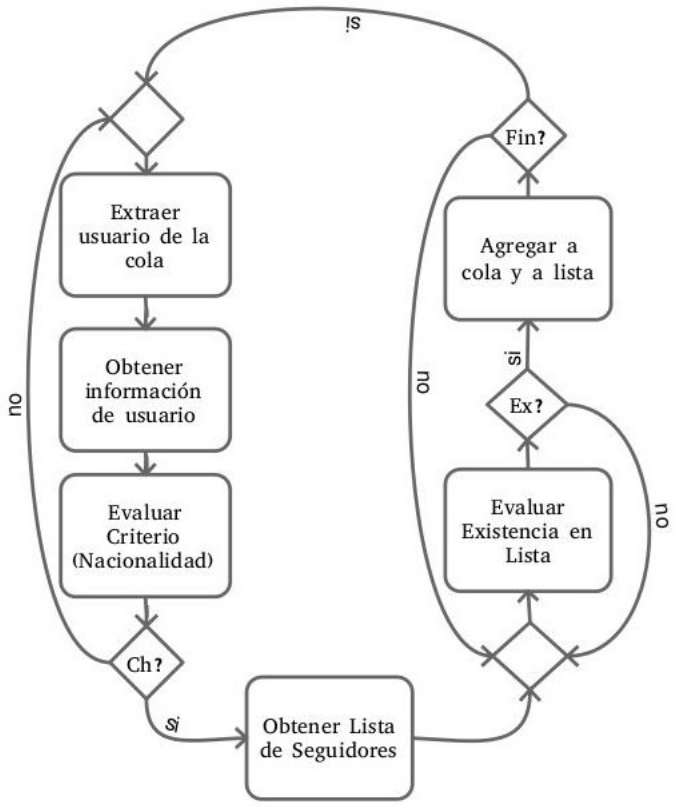

Figura 4.4: Crawler de Usuarios
Fuente: Elaboración Propia

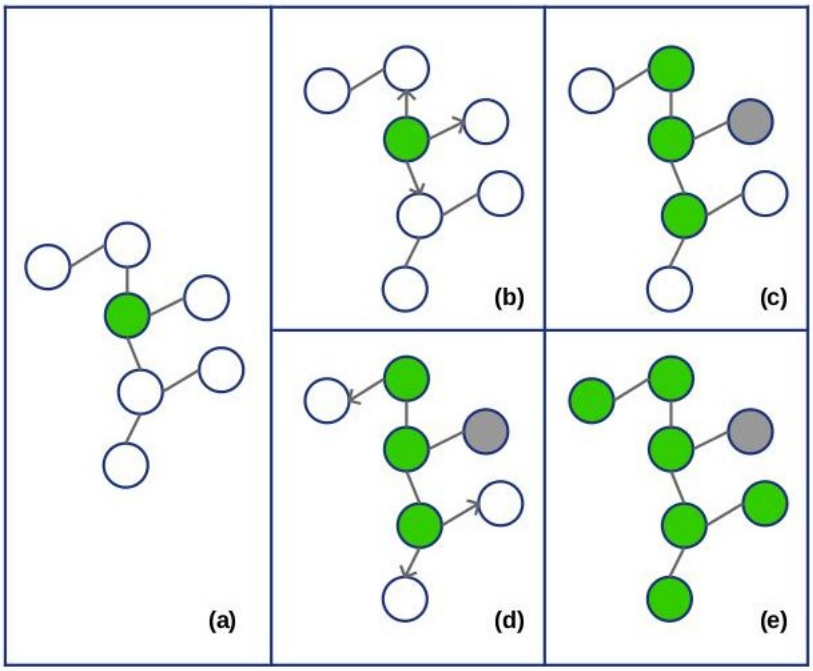

Figura 4.5: Ejemplo de Iteración
Fuente: Elaboración Propia

tropolitana, los ríos, entre otras)

- Nivel comunal: Las comunas de Chile que cuenten con 30.000 habitantes o más para el CENSO 2012 (por ejemplo, Panguipulli).
- Ubicación geo-referenciada: Campos consignados de la forma de tupla (latitud, longitud) correspondiente a Chile.
- Otras: Opciones alternativas como "shile", "stgo", entre otras.

Es importante mencionar que existe un porcentaje de usuarios de *Twitter* que bloquean el acceso a sus *tweets*. Es una opción habilitada en la configuración del usuario, pero no es una opción ampliamente usada. Sin embargo, para efectos del diseño, no serán considerados los usuarios con esta condición, debido a que no agregan valor significativo a los indicadores. Esta elección significa agregar un criterio más al "Crawler Focalizado" de usuarios.

Una vez establecida la base de usuarios se procederá a extraer el conjunto de *tweets* publicado por cada uno. La consulta se hace incluyendo la ID del usuario, sin embargo sólo es posible acceder hasta los 3200 *tweets* más recientes. Esta limitación es imposible de eludir, por lo tanto se asume como dato. Por otro lado, es imperioso filtrar los *tweets* desde el principio bajo algún criterio, debido a que la mayoría de los tweets no aportan a los indicadores y su inmensa cantidad implican alto costo computacional de almacenamiento. Por esto se guardarán sólo aquellos *tweets* relacionados con marihuana.

El modo de clasificar los *tweets* relacionados con marihuana consistirá en identificar palabras clave. El listado de palabras clave tendrá como origen a tres fuentes diferentes: conocimiento experto, bibliografía y una encuesta de uso de palabras actuales. La primera es facilitada por el cliente del proyecto, la segunda se compone por [70] y [22], y la tercera, una encuesta realizada por cuenta propia. Las primeras dos fuentes contienen palabras relacionadas con la marihuana y su uso. La última quiere responder a la pregunta: "¿Cuál o cuáles son los términos o expresiones que usted utiliza o conoce para referirse a la marihuana o sus formas de uso?", seleccionando las palabras más repetidas.

El conjunto total de palabras será filtrado para confirmar su uso en *Twitter*, con el objetivo de desambiguar el contexto de uso. Para hacer esto se extraerán desde *La Gorda* todos los *tweets* que contengan las palabras. Serán aplicados los siguientes algoritmos de pre-procesamiento de texto:

- Tokenización
- Remoción de Palabras Vacías (*stopwords*)
- Eliminación de emoticones y emojis (por ejemplo, ":)")
- Eliminación de elementos de Twitter (por ejemplo, retweets)

Será aplicado *Topic Modeling* con el objetivo de identificar contextos diferentes de uso de las palabras y verificar su empleo relacionado con marihuana. Los *tweets* por palabra clave serán divididos en varios grupos, determinados automáticamente por el modelo, de los cuales se identificarán manualmente aquellos con la característica buscada y sus palabras en común. Luego, las palabras ambiguas serán filtradas, considerando sólo aquellos *tweets* que contengan la cadena de caracteres "fum". Esta regla parece ser muy restrictiva, pero un análisis exploratorio la arrojó como la palabra común más utilizada en el contexto. La tarea de incorporar otras reglas resulta ser muy consumidora de tiempo, ya que es necesario comparar los resultados de cada palabra. En caso de no encontrar

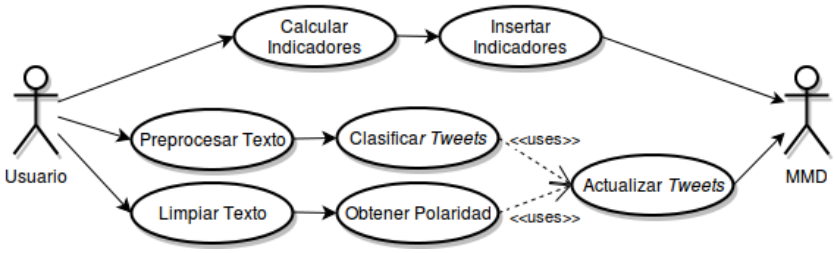

Figura 4.6: Casos de uso de MIT
Fuente: Elaboración Propia

tweets para alguna palabra clave, esta será descartada automáticamente.

4.4.2. Inteligencia de *Tweets*

El Módulo de Inteligencia de *Tweets* (MIT) será responsable, en una primera etapa, del tratamiento de textos y en la segunda, de la respectiva clasificación y el cálculo de polaridad. La primera etapa es imprescindible para el desarrollo de la segunda, porque es imposible aplicar las técnicas de clasificación de texto sin una estructura que permita reconocer parámetros. En esta sección serán abordadas las técnicas de procesamiento de texto que permitan representar un conjunto de documentos en forma matricial. Por otro lado, serán señaladas las clasificaciones necesarias por los indicadores y los algoritmos que serán aplicados para obtener dichas clasificaciones. Además se mencionará el uso de una aplicación de *Opinion Mining* para el cálculo de polaridad de los textos. En la figura 4.6 se muestran todos los casos de uso del módulo.

El preprocesamiento busca una representación matricial del conjunto de documentos. En la práctica, el límite de 140 caracteres de los *tweets* y tipo de clasificación condicionan la serie de las transformaciones y la naturaleza de cada celda. Tomando en cuenta esto, son consideradas las siguientes etapas en el pre-procesamiento:

- Tokenización
- Normalización
- Eliminación de caracteres especiales
- Eliminación de elementos de Twitter (por ejemplo, RT)
- *Stemming*
- Formulación de n-gramas
- Remoción de Palabras Vacías (*stopwords*)
- Representación de documentos (vectores)

Las clasificaciones son del tipo binario, es decir, el texto es evaluado por su pertenencia a cierta categoría. Como se ha mencionado antes, estas categorías son el consumo de marihuana, la mención de políticas de control de la droga y la venta de marihuana. Cada categoría se evaluará independientemente de las otras. Se utilizarán algunos de los algoritmos que han mostrado mejor rendimiento en clasificación de textos según la literatura ([21] y [38]), como también otros alternativos. Algunos son:

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

En este módulo será aplicado el cálculo de polaridad de los tweets, mediante una API de *Opinion Mining* perteneciente al proyecto OpinionZoom, llamada PAPI. Esta medida será aplicada a cada *tweet*, haciendo posible la adquisición de información agregada de los sentimientos de las opiniones sobre marihuana. Antes de que cada *tweet* sea procesado, será tratado previamente para eliminar ruido que interfiera con los algoritmos de la API. El ruido está en la forma de elementos de *Twitter* y otras caracteres especiales no utilizados en la comunicación escrita convencional. Otros pasos no son necesarios porque la API tiene su propia forma de preprocesar los textos.

Una vez que son calculados todos los atributos correspondientes a un *tweet* se procede a estimar los valores a agregados, es decir, los indicadores para el conjunto de *tweets*. En resumen, se deben calcular los siguientes datos diariamente:

- Polaridad promedio positiva, negativa y total.
- Número de *tweets* positivos, negativos y neutros.
- Identificar la mención del tema, consumo, políticas de control y venta de marihuana.
- Segmentar por *keyword*.

Finalmente, se procederá a guardar el conjunto de valores diarios en la base de datos. Esta medida debe realizarse debido a la de evolución constante de *Twitter* y por la estabilidad requerida de los datos. Se desea trazar una evolución de los datos a través del tiempo, pero esto es imposible si los datos de *Twitter* están constantemente cambiando. Esto produciría que la forma de la linea de tiempo de los indicadores también cambiara. Por esta razón, la actualización diaria de los *tweets* representa una foto en el tiempo del conjunto de datos y la base de datos se encargará de almacenar el álbum de fotos.

4.4.3. Inteligencia de Usuarios

El Módulo de Inteligencia de Usuarios (MIU) se hará cargo de determinar el conjunto de datos que caracteriza al usuario, su consumo y su opinión sobre la marihuana. En primer lugar será mencionada la fase de pre-procesamiento de texto, ya que también se utilizará un conjunto de *tweets* del usuario para calcular su edad. Luego, se tratará la utilización de las relaciones entre los usuarios y su uso para calcular las métricas de Análisis de Redes Sociales. Por otro lado, se mencionará el conjunto de variables usadas para predecir el consumo. En la parte final, se precisarán los métodos de clasificación de edad y clasificación de consumo. En resumen, se verán todos lo casos de uso del módulo, representados en la Figura 4.7.

La segmentación por edad de los usuarios es una etapa clave en el desarrollo de la aplicación. Esta medida permitirá determinar la presencia de los distintos rangos etarios en la red de usuarios chilenos de *Twitter*, apoyar la clasificación de consumo y apreciar de forma más detallada el consumo de marihuana. El proceso se guiará por el trabajo realizado por D. Nguyen en [65], utilizando

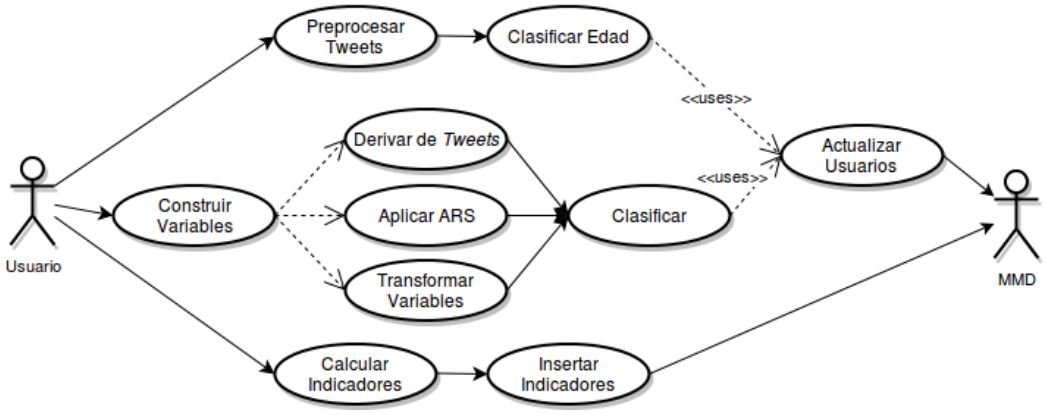

Figura 4.7: Casos de uso de MIU
Fuente: Elaboración Propia

el conjunto de *tweets* de cada usuario. Por esto, es necesario realizar el mismo tratamiento de texto que el descrito en la sección anterior, con excepción de n-gramas, ya que utiliza solo uni-gramas.

Teniendo presente lo anterior, es necesario diseñar una mezcla de atributos que permitan el mejor rendimiento de los clasificadores, teniendo en cuenta el uso de sólo datos obtenidos de *Twitter*. Con este objetivo en mente, se usará el siguiente conjunto de atributos para clasificar el consumo de marihuana por usuario:

- Número de menciones de marihuana
- Número de menciones de consumo
- Número de menciones de políticas
- Métricas de polaridad del usuario
- 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

Algunos medidas de Análisis de Redes Sociales fueron descartadas por su alto costo computacional, imposibilidad de cálculo y porque la literatura mostró que no tenían influencia significativa en el consumo de marihuana. Además fue agregada una medida de polaridad del vecindario, para reproducir una cierta especie de norma social. A continuación se muestra el detalle de las fórmulas y algunas definiciones:

- Conjunto de vértices (usuarios): $V = \{1, 2, ..., n\}$
- Conjunto de seguidores del usuario k: $F_k = \{1, 2, ..., n_k\}$
- Conjunto de amigos del usuario k: $A_k = \{1, 2, ..., m_k\}$
- Vecindario del usuario k: $N_k = F_k \cup \{k\}$
- Densidad de vecindario:

$$\mathrm{de}nsity_k = \frac{2 * (|F_k| + \sum_{\mathrm{i} \in F_k} \sum_{j \in F_\mathrm{i}} b_j^{N_k})}{|N_k| * (|N_k - 1|)}$$

  donde

$$b_j^{N_k} = \begin{cases} 1 & \text{si } j \in N_k \\ 0 & \text{si no} \end{cases}$$

- *Indegree*:

$$\mathrm{in}_k = |F_k|$$

- *Outdegree*:

$$out_k = |A_k|$$

- *Reach Centrality*:

$$reach_k = \frac{|\bigcup_{j \in F_k} F_j \cup F_j - \{k\}|}{|V|}$$

- Conjunto de usuarios que mencionan consumo: $M = \{1, 2, ..., w\}$
- Amigos consumidores:

$$m_k = \frac{\sum_{j \in A_k} c_j^M}{|A_k|}$$

  donde

$$c_j^M = \begin{cases} 1 & \text{si } j \in M \\ 0 & \text{si no} \end{cases}$$

- Polaridad de vecindario:

$$p_k = \frac{\sum_{j \in A_k} \mathrm{d}_j}{|A_k|}$$

  donde d_j es la polaridad del usuario j
- Distancia:

$$\mathrm{dis}t_k = \begin{cases} 1 & \text{si } A_k \cap M \neq \emptyset \\ 2 & \text{si } A_k \cap M = \emptyset \wedge \bigcup_{j \in A_k} A_j \cap M \neq \emptyset \\ 3 & \text{si no} \end{cases}$$

- Nominaciones externas:

$$outnom_k = friends_k - out_k$$

La mayoría de los atributos son calculados en base a los datos obtenidos para cada *tweet*. En consecuencia, se necesitan algunas transformaciones para traducirlas a datos del usuario. Los atributos que requieren mayor procesamiento son aquellos asociados con las métricas de Análisis de

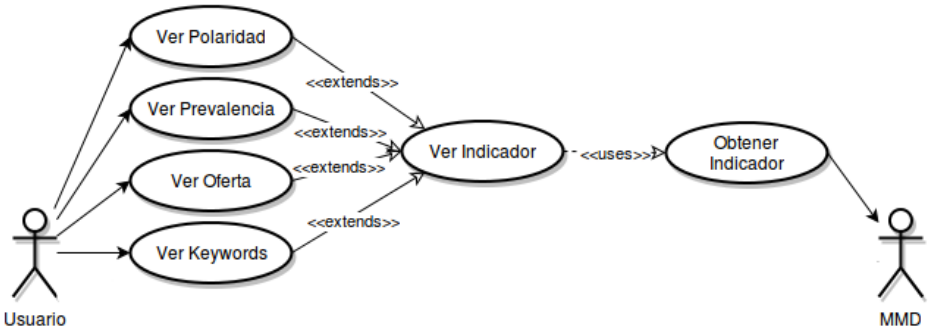

Figura 4.8: Casos de uso de MVR
Fuente: Elaboración Propia

Redes Sociales. Las métricas, que fueron detalladas en el capítulo anterior, harán uso de las relaciones entre usuarios y algunos de sus datos individuales. Para algunas métricas se utilizarán los seguidores y para otras los amigos.

La clasificación se aplicará haciendo uso de algoritmos diferentes para la edad y el consumo. En el primero se replicará el uso de regresiones lineales, tal como se hizo en [65], y se explorarán otros algoritmos. Se calculará primero la edad y luego se hará la categorización por rango. Según Wolpert en [94], no existe tal cosa como el mejor clasificador. Los algoritmos difieren en rendimiento para cada problema en particular. Por lo tanto, para el segundo problema de clasificación se evaluará el uso de varios algoritmos clásicos, como los siguientes:

- *Support Vector Machines*
- Redes Neuronales
- *Naive Bayes*

Por otra parte, tal como se estableció en la sección anterior, se procederá a guardar un registro de todas las actualizaciones de los indicadores. Pero a diferencia de los *tweets*, los datos serán actualizados mensualmente. En cada actualización se realizarán cada uno de los pasos, con excepción de la evaluación de los algoritmos, ya que la elección será tomada una vez, en función de los rendimientos de cada uno.

4.4.4. Visualización de Resultados

El Módulo de Visualización de Resultados (MVR) se hará cargo de presentar los resultados de una manera amigable y comprensible para el usuario final. Con el objetivo en facilitar el acceso a los datos, se procederá a implementar una aplicación web. Los datos estarán cambiando frecuentemente, por lo tanto es necesaria una herramienta que permita visualizar estos cambios de manera continua. Este módulo considera todos los elementos para visualizar los indicadores e información importante para el usuario. Tal información es representada por los casos de uso que son mostrados en la Figura 4.8.

En base a lo observado en la Figura 4.8 se ve que la motivación principal es mostrar indicadores.

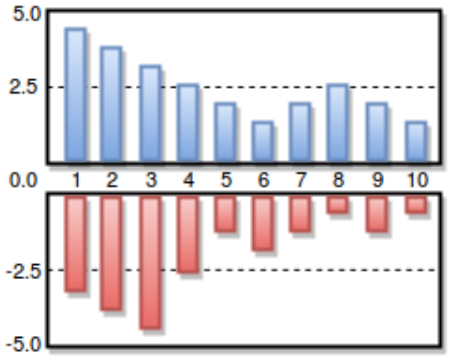

Figura 4.9: Gráfico de polaridad
Fuente: Elaboración Propia

Pero deben ser exhibidos de la forma que favorezcan el entendimiento de la información por parte del cliente. La forma de mostrar los datos será determinada por el tipo de indicador que se quiere apreciar. Los cuales son enumerados a continuación:

1. Polaridad: La manera principal de mostrar la polaridad será mediante gráficos como el de la Figura 4.9. Las barras en la parte superior muestran el promedio de la polaridad positiva, mientras que las barras en la parte inferior muestran el promedio de la contraparte negativa. Para efectos de la aplicación, el eje horizontal podrá ser definido por divisiones temporales (día, semana, mes, año) y por palabras claves. La segunda manera de mostrar polaridad será como en la Figura 4.10, donde se muestra la polaridad promedio para un conjunto de elementos con polaridad asociada. La tercera forma será mediante un gráfico como en la Figura 4.11, donde el eje vertical muestra el número total de elementos y es descompuesto por el número de elementos positivos, neutros y negativos.

2. Prevalencia: La definición de prevalencia ayuda a definir la forma de mostrar los datos. Debido a que la prevalencia representa el porcentaje de usuarios de la población que presentan consumo, ayuda cualquier gráfico que muestre una serie de tiempo de porcentajes. Esta misma forma se aplicará para otros indicadores como la del promedio de "amigos" consumidores y porcentaje de usuarios mencionando el consumo de marihuana. El gráfico tipo se muestra en la Figura 4.12.

3. Oferta: Estos datos tienen el objetivo de mostrar la cantidad de cuentas asociadas a venta de marihuana, así como el porcentaje de ellas. Por lo tanto, serán utilizados gráficos como los mostrados en las Figuras 4.11 y 4.12.

4. Palabras Clave: Se considera que la forma de presentar una gran cantidad de palabras, de forma de apreciar su importancia relativa, es tal como se muestra en la Figura 4.13. Es una nube de palabras elaborada por el sitio *TagCrowd* desde un archivo con palabras utilizadas actualmente para referirse a la marihuana (encuesta). Además se mostrará el historial de frecuencia y la distribución de cada palabra en base a su utilización en *tweets* de consumo, políticas y venta. Para esto, de nuevo son útiles el tipo de gráficos mostrados en las Figuras 4.11 y 4.12.

Lo mencionado hasta ahora forma parte del contenido de la aplicación web. En lo que queda de

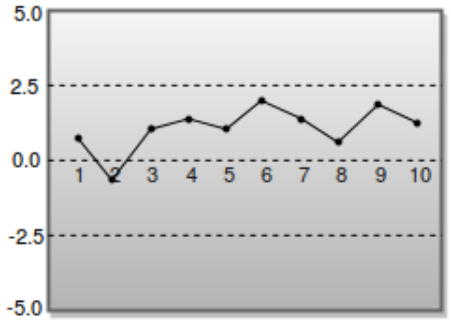

Figura 4.10: Gráfico de polaridad promedio
Fuente: Elaboración Propia

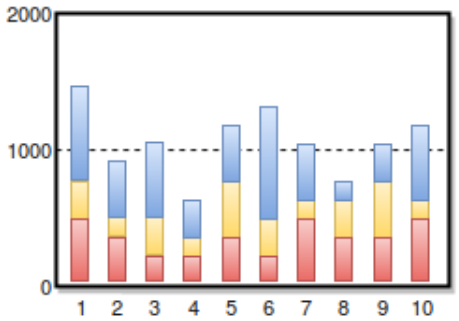

Figura 4.11: Gráfico de total de polaridad
Fuente: Elaboración Propia

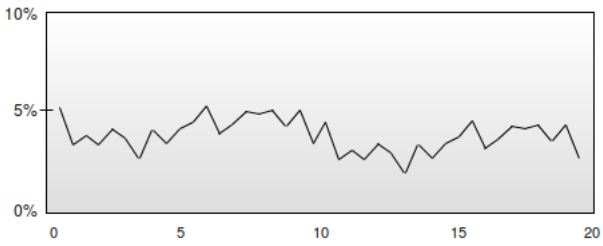

Figura 4.12: Gráfico de prevalencia
Fuente: Elaboración Propia

Figura 4.13: Nube de palabras de encuesta
Fuente: Sitio TagCrowd.com

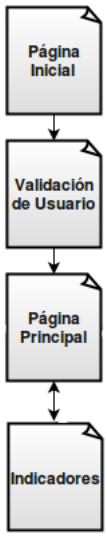

Figura 4.14: Estructura Navegacional
Fuente: Elaboración propia

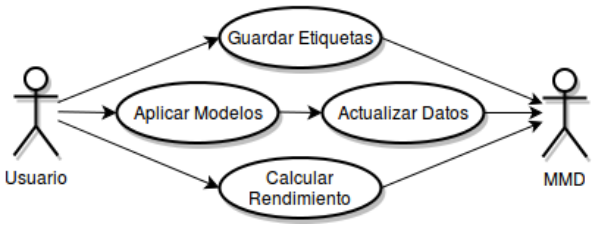

Figura 4.15: Casos de Uso de MER
Fuente: Elaboración propia

sección se mostrará el diseño de la estructura navegacional. En efecto, en la Figura 4.14 se puede apreciar la estructura navegacional que cuenta con los siguientes partes:

- Página de inicio: página con un mensaje de bienvenida.
- Validación de usuario: ingreso de usuario con nombre y clave de acceso.
- Página principal: se muestran los principales indicadores y un menú con los enlaces a las demás páginas.
- Indicadores: página de visualización adaptable a cada indicador.

4.4.5. Evaluación de Rendimiento

El Módulo de Evaluación de Rendimiento (MER) determinará el grado de certeza con que son calculados los datos para su consiguiente uso en los indicadores. La Figura 4.15 muestra los casos de uso que resumen las funcionalidades del módulo. A continuación se describe cada una:

- Guardar etiquetas: esta funcionalidad identifica las etiquetas en los archivos clasificados manualmente y las empareja con su caso correspondiente. En el caso de ser *tweet*, se reconocen las etiquetas de mención, consumo, políticas de control y venta de marihuana. En el caso de

ser usuario, identifica las etiquetas manuales para la edad y el consumo de marihuana del usuario. Después de hacer esto, realiza la parte inicial del proceso de insertado de etiquetas en la base de datos.

- Aplicación de modelos: una vez que son determinados los parámetros para cada modelo, esta función se encarga de recolectar los casos de estudio etiquetados manualmente y aplicarles los modelos evaluados. Para cada uno de los modelos y para cada uno de los clasificadores se lleva a cabo la tarea de clasificar los casos y posteriormente almacenarlos en la base de datos.
- Calcular rendimiento: esta parte se encarga de la tarea final del módulo de calcular las métricas de evaluación de rendimiento para cada modelo. Determina el valor de cada métrica para todos los problemas de clasificación y entrega la comparación que permitirá la elección del conjunto de los modelos a emplear en la aplicación.

4.4.6. Mantenimiento de Datos

El Módulo de Mantenimiento de Datos (MMD) administrará todas las operaciones de la base de datos. La existencia de este módulo es crucial para la aplicación, ya que interactúa con todos lo módulos, a excepción del MMD mismo. En síntesis desempeñará las tareas de leer, actualizar e insertar datos cuando corresponda. Además de proveer la capa de almacenamiento de datos y su lógica. Todos los casos de uso para este módulo están separados por los módulos con los cuales interactúa y son mostrados en la Figura 4.16. En lo que queda de capítulo se procederá a describir la lógica de la capa de datos y las interacciones del MMD.

Para modelar los datos pertenecientes al sistema de información se hará uso de un modelo Entidad-Relación (E-R). El modelo soporta una colección acotada de entidades, pero estas entidades contienen toda la información importante para toda la aplicación. La estructura está dada por los datos que son utilizados como entrada para los indicadores. Estos elementos son usuarios, *tweets* y relaciones entre usuarios. Aparte de estos, es agregado un elemento que sostiene la lógica de recolección de datos: las palabras clave. El esquema E-R se muestra en la Figura 4.17. La composición del modelo E-R se describe a continuación.

- Usuario: los usuarios en sí representan el origen de todo el contenido de *Twitter*. Esta es la razón principal para tenerlos en la base de datos. Para efectos de los indicadores tendrán como conjunto de atributos, uno formado por: mención de consumo, mención de venta, un índice M calculado por los clasificadores, la predicción de consumo y la frecuencia de consumo. También tiene otros atributos que ya han sido nombrados a lo largo del capítulo.
- *Tweet*: estos son la base de toda la información, la base de todo el conocimiento obtenido desde los usuarios. Por lo tanto, deben mantener una relación directa con ellos en la base de datos. Además poseen atributos en sí como la polaridad, mención de venta, mención de políticas y mención de consumo de marihuana. Además contiene la información en sí del *tweet*.
- *Keyword* (Palabra Clave): esta entidad tiene un registro de las palabras relacionadas con marihuana que son encontradas en cada *tweet*.
- Relación: aunque el nombre es sugerente, representa la relación entre usuarios en el modelo. Es excluida cualquier relación con usuarios fuera de la base de datos.

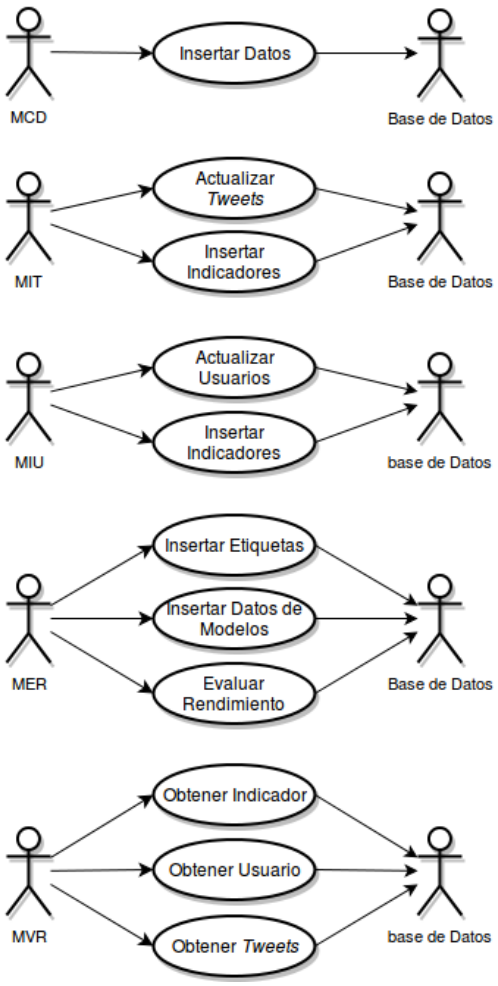

Figura 4.16: Casos de Uso de MMD
Fuente: Elaboración propia

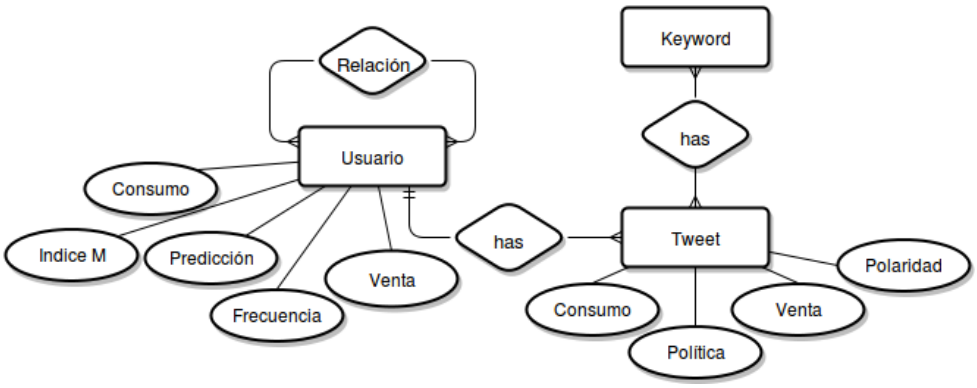

Figura 4.17: Diagrama E-R
Fuente: Elaboración propia

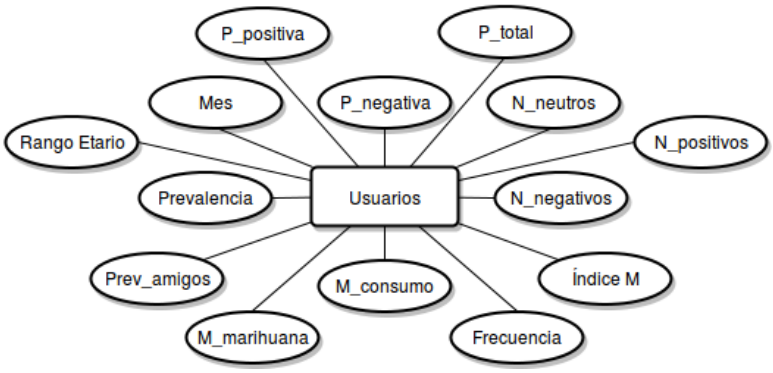

Figura 4.18: Entidad Usuarios
Fuente: Elaboración propia

Es requerido otro modelo para representar la lógica de los datos para la evaluación de rendimiento. No será presentado en forma gráfica porque es casi una réplica del descrito en la parte anterior. Pero cabe mencionar modificaciones como son:

- Entidades: sólo serán modelados las entidades "usuario" y "*tweet*", ya que las demás son innecesarias para el propósito.
- Relaciones: se conservan las relaciones entre las dos entidades consideradas y es eliminada la relación entre usuarios "relación", ya que es incorporada en forma de atributos en la entidad "usuario".
- Atributos: son agregados conjuntos de atributos a las dos entidades. Son réplicas para las clasificaciones, ya que necesitan la parte de etiquetado manual y la clasificación de los modelos. Por esto, es necesario multiplicar el número columnas de etiqueta por cada algoritmo considerado en la evaluación.

Por otra parte, los módulos de Inteligencia de Usuario, Inteligencia de *Tweets* y Visualización de Resultados hacen necesaria la adición de otros elementos a la lógica de la base de datos. Estos elementos sostienen el seguimiento de los indicadores en el tiempo y el acceso restringido a los mismos. Cada entidad es independiente de la otra, por lo que son mostrados por separado en las Figuras 4.18, 4.19, 4.20 y 4.21. En las primeras tres figuras los atributos para cada entidad contienen información relevante para la composición de los indicadores y son más o menos explicativos, por lo que no se explicarán en detalle. La última figura no será descrita por su simplicidad.

Finalmente, serán descritas las interacciones con cada módulo de la aplicación. En general son utilizadas casi todas las operaciones posibles en una base de datos:

- MCD: El módulo de recolección de datos requiere la inserción de los datos obtenidos desde *Twitter*. Estos son conjuntos de usuarios y *tweets* por separado. Los datos deben ser ingresados en las entidades "usuario" y "*tweet*".
- MIT: Este módulo requiere que se actualicen los datos que dan como resultado su proceso interior, e insertar los indicadores que se construyen con esos datos. La primera tarea consiste en hacer actualizaciones de la entidad "*tweet*" y la segunda, inserciones en la entidad "*tweets*".
- MIU: La tareas de este módulo hacen uso de las mismas operaciones que el MIT, pero con

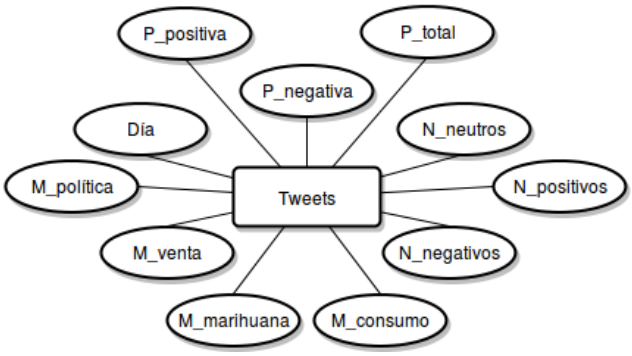

Figura 4.19: Entidad *Tweets*
Fuente: Elaboración propia

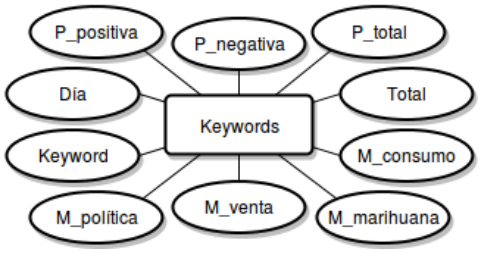

Figura 4.20: Entidad Keywords
Fuente: Elaboración propia

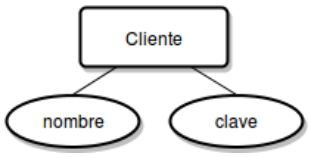

Figura 4.21: Entidad Cliente
Fuente: Elaboración propia

respecto a las entidades “usuario” y “usuarios”.

- MER: Este módulo hace uso de las entidades “usuario” y “*tweet*”, pero del modelo E-R correspondiente a la evaluación de rendimiento. Se separa por la tarea de inserción de etiquetas y resultados de los clasificadores, y por la tarea de obtener los datos ingresados anteriormente para calcular las medidas de rendimiento.
- MVR: Este módulo utiliza las entidades “usuarios”, “*tweets*” y “*keywords*” para obtener los indicadores que serán mostrados en la aplicación web.

Capítulo 5

Implementación

Luego del capítulo anterior, se tienen los planos generales para la construcción de la aplicación, los cuales son consignados desde un nivel más abstracto. En este capítulo es hora de materializar las ideas y detallar los vehículos utilizados para dicha tarea. Esta etapa es sumamente crítica ya que permite apreciar la factibilidad de las ideas y condicionará los resultados que serán presentados en el siguiente capítulo.

La primera parte de este capítulo será dedicada a describir las herramientas utilizadas, debido a que estas determinan el resultado final a partir de sus libertades y limitaciones. La segunda parte se destinará a detallar la forma en que fueron seleccionadas las palabras clave, relacionadas con marihuana. Luego, se especificará cómo fue realizado el etiquetado de *tweets* y usuarios. La última parte seguirá con el orden planteado en el capítulo anterior, es decir, la implementación será descrita desde el punto de vista de cada módulo de la aplicación. Esto facilitará la comprensión de la aplicación como la unión de un conjunto de módulos que son implementados por separado, pero que juntos permiten el funcionamiento de la aplicación general.

5.1. Herramientas de Desarrollo

Las herramientas utilizadas serán descritas ordenadamente según su propósito y origen. Primero será mencionado el nivel de software más amplio, es decir, el sistema operativo. Luego será descrito el lenguaje de programación y librerías que facilitaron el desarrollo. Finalmente, serán detallados los sistemas de administración de bases de datos y el marco empleado para el desarrollo de la aplicación web.

- **Ubuntu:** es un sistema operativo basado en GNU/Linux, que se distribuye como software libre. Cuenta con una gran variedad de versiones, y ediciones especiales para servidores, nubes OpenStack y dispositivos móviles. Todas comparten infraestructura y software, destacando a la plataforma por su gran escalabilidad. Durante el desarrollo del proyecto fue utilizada la versión de escritorio 14.04.4 LTS y su edición para servidores. Esta es la más reciente de las versiones estables que reciben soporte por parte de la comunidad mundial de desarrolladores durante periodos de 5 años. Según su sitio oficial, es el sistema operativo de software

libre más popular entre desarrolladores, debido a su versatilidad, confianza, actualización constante, y amplias librerías de desarrollo.

- **Java:** es un lenguaje de programación compilado y plataforma tecnológica que incluye un conjunto de herramientas comerciales y de software abierto, un extenso ecosistema de experimentados desarrolladores, y una plataforma estandarizada de desarrollo de aplicaciones. Es el lenguaje de programación más popular según el índice TIOBE. Su popularidad se sostiene en dos características: no depende del hardware y del sistema operativo, ya que sólo es necesaria la existencia de una Máquina Virtual de JAVA, y es un lenguaje orientado a objetos, una cualidad de sumo interés por parte de los desarrolladores.
- **Twitter4j:** es una librería de Java no oficial para operar con la API de Twitter. Ofrece recursos para interactuar con la API Streaming y la API Rest, las cuales proveen información sincronizada e histórica, respectivamente. Además contiene recursos para el manejo de credenciales, necesarias para acceder a la mayoría de datos contenidos en Twitter, tales como información de usuarios, *tweets*, relaciones de seguimiento, entre otros.
- **Librerías de Procesamiento de Texto:** en una primera instancia fueron unidas algunas librerías independientes de procesamiento de texto. JTextPro es un conjunto de herramientas de procesamiento de texto basadas en Java que incluye separadores de oraciones, tokenización de palabras, etiquetado Part-of-speech y fragmentación de frases. Además fue utilizado un algoritmo libre de *Stemming* basado en el algoritmo "*Snowball*" de Martin Porter.
- **LDAGibbs:** es una implementación de Java de *Latent Dirichlet Allocation* (LDA) que usa la técnica de *Gibbs Sampling* para estimación de parámetros e inferencia. Este ejemplo de *Topic Modeling* permite la obtención de todos los parámetros calculados por LDA desde un archivo que contiene los textos a analizar. Los parámetros del modelo deben ser ingresados manualmente.
- **Weka:** según su sitio oficial, Weka es distribuido como un software libre. Está constituido como una colección de algoritmos de *Machine Learning* para realizar minería de datos. Los algoritmos pueden ser aplicados directamente mediante su propia interfaz o pueden ser llamados desde código Java. Weka contiene instrumentos para realizar pre-procesamiento, clasificación, regresión, *clustering*, reglas de asociación, y visualización.
- **Gephi:** es un software libre que permite la visualización y exploración de todo tipo de grafos y redes. Esta constituida como una librería estándar de Java que contiene módulos separados para la estructura de grafos, algoritmos de distribución de grafos, filtros, entre otros. Además contiene algoritmos típicos para el Análisis de Redes Sociales (SNA, por sus siglas en inglés).
- **FreeLing:** consiste en un librería escrita en C++ que facilita servicios de análisis de lenguaje. Provee APIs para la interacción de la librería con otros lenguajes, como por ejemplo Java. *Freeling* ofrece servicios de tokenización, separación de oraciones, análisis morfológico, etiquetado PoS, entre otros. Actualmente soporta varios lenguajes, incluyendo el español.
- **PostgreSQL:** según su sitio oficial, PostgreSQL es un sistema de gestión de bases de datos relacional orientado a objetos, distribuido como software libre. Posee una gran reputación por su confianza, integridad de los datos, y exactitud. Opera en la mayoría de sistemas operativos, y además de trabajar con tipos básicos de datos, soporta imágenes, sonidos y videos. También tiene interfaces de programación nativas para C/C++, Java, Perl, Ruby, Python, entre otros.
- **JSON:** es un formato liviano de intercambio de datos que debe su nombre a las siglas en inglés de *JavaScript Object Notation*. Está basado en un subconjunto del lenguaje de programación *JavaScript* y es un formato de texto que es completamente lenguaje-independiente.

JSON está construido en base a una colección de pares de nombres y valores, y una lista ordenada de valores. Es soportado virtualmente por todos los lenguajes de programación.

- **ArangoDB:** es una base de datos multi-modelo NoSQL , distribuida como software libre con modelos de datos flexibles para documentos, grafos y pares llave-valor. Implementa un lenguaje de consulta parecido al SQL y almacena los datos en forma de documentos con formato JSON.
- **CodeIgniter:** es un marco de desarrollo de sitios y aplicaciones web desarrollado en PHP, distribuido como software libre. CodeIgniter promueve el uso de modelo MVC y facilita una extensa librería que cubre muchas de las necesidades del desarrollo web. Actualmente es uno de los marcos de desarrollo más populares de PHP.

5.2. Selección de Palabras Clave

La selección de palabras clave consiste, a grandes rasgos, en un proceso de evaluación y filtrado de un conjunto de palabras iniciales. Por lo tanto, es lógico pensar que es necesario fijar qué es lo que será evaluado y cómo. En primer lugar, para cada palabra se requería extraer de La Gorda un conjunto de textos asociados a cada palabra. Luego, realizar tratamiento de texto para obtener información de este. Finalmente, evaluar la utilización de la palabra. Para hacer esto se implementaron varias clases en Java y se utilizaron otras librerías:

- Extractor de textos: se implementó una clase que permitió buscar palabras en una base de datos local. Para hacer esto se recurrió a un archivo con una lista de palabras y se buscó todos los textos que contuvieran cada una de ellas. Los textos fueron guardados en nuevos archivos esperando su respectivo tratamiento.
- Pre-procesamiento de texto: fueron utilizadas las librerías Java de pre-procesamiento de texto para ordenar los textos, quitar elementos innecesarios, tokenizar, quitar *stopwords*, normalizar y realizar *Stemming*. Aunque en la práctica este último no fue requerido.
- Evaluación de Contexto: para las palabras con muchos textos asociados se aplicó *Topic Modeling* para separar los documentos en grupos que compartieran características, en este caso, los grupos (tópicos) son representados por una distribución de palabras. En base a esto se separaron las palabras de contextos ambiguos y otros netamente relacionados con marihuana. En caso de palabras con pocos textos se evaluó manualmente.
- Aplicación del Criterio "fum": se escribió una clase sólo para evaluar la presencia de la cadena de caracteres "fum" y calcular estadísticas. Este punto resulta crucial para determinar la utilización de palabras de contexto ambiguo.

5.3. Etiquetado

El etiquetado de casos de *tweets* y usuarios se realizó mediante la implementación de dos sitios web, uno para cada tipo de etiquetado. Se utilizó el marco de desarrollo CodeIgniter, apegándose a la arquitectura Modelo-Vista-Controlador y realizando escrituras por medio del gestor de bases de datos relacionales PostgreSQL.

5.3.1. Etiquetado de *Tweets*

El etiquetado de *tweets* en sus respectivas categorías se hizo mediante manejo de sesiones. Los etiquetadores ingresaban como usuarios identificados y se mostraban ante ellos los casos que les correspondía clasificar. El flujo de *tweets* era completamente lineal, ya que los casos aparecían uno después de otro hasta que se completara el proceso. Las etiquetas eran guardadas automáticamente, por lo que no era necesario realizar el proceso completo en una sola sesión. También existía la opción de revisar casos clasificados anteriormente.

5.3.2. Etiquetado de Usuarios

El etiquetado consistía en una encuesta directa a los usuarios de *Twitter*, por ende requería implementar dos elementos: el mecanismo de esparcimiento y la propia encuesta.

- Mecanismo de Esparcimiento: fue implementado un algoritmo que utiliza la librería *Twitter4j* para publicar *tweets* a una lista de usuarios, mencionando la persona a la que está dirigido y direccionando a la ubicación del sitio web de la encuesta. La clase también hace manejo de credenciales, el elemento crítico para interactuar con *Twitter*.
- Sitio Web de Encuesta: el sitio utiliza un flujo sumamente lineal para completar las cuatro preguntas requeridas. Primero se visualiza la página de bienvenida, luego se muestra la autorización de un consentimiento informado, después se procede a mostrar las tres preguntas relacionadas con el estudio y finalmente, se requiere el ingreso de la cuenta de *Twitter*. Todo esto es guardado en la base de datos. Algunos campos fueron implementados con validación especial, para ello se mostraban los errores asociados.

5.4. Implementación de la Aplicación

Antes de detallar la forma en que los módulos fueron implementados, cabe destacar que la aplicación fue construida en torno al lenguaje de programación Java. Las librerías en otros lenguajes y bases de datos son llamadas mediante librerías de Java. El Módulo de Visualización de Resultados es el único que escapa de esta regla, ya que realiza consultas directas a la base de datos. A continuación son descritos cada uno a los módulos, mencionando las clases principales que conforman la estructura de la aplicación. Son omitidas algunas componentes que facilitan el funcionamiento, pero que no son imprescindibles o son ampliamente conocidas.

5.4.1. Módulo de Recolección de Datos

Este módulo es el encargado de interactuar con *Twitter*, por lo tanto hace uso intensivo de la librería construida para esa tarea, *Twitter4j*. La librería en sí permite obtener toda la información pública de los usuarios de *Twitter*, además de realizar acciones de escritura para el usuario asociado a la credencial en uso. La solución a las restricciones impuestas por la *API Rest* de *Twitter*, y la

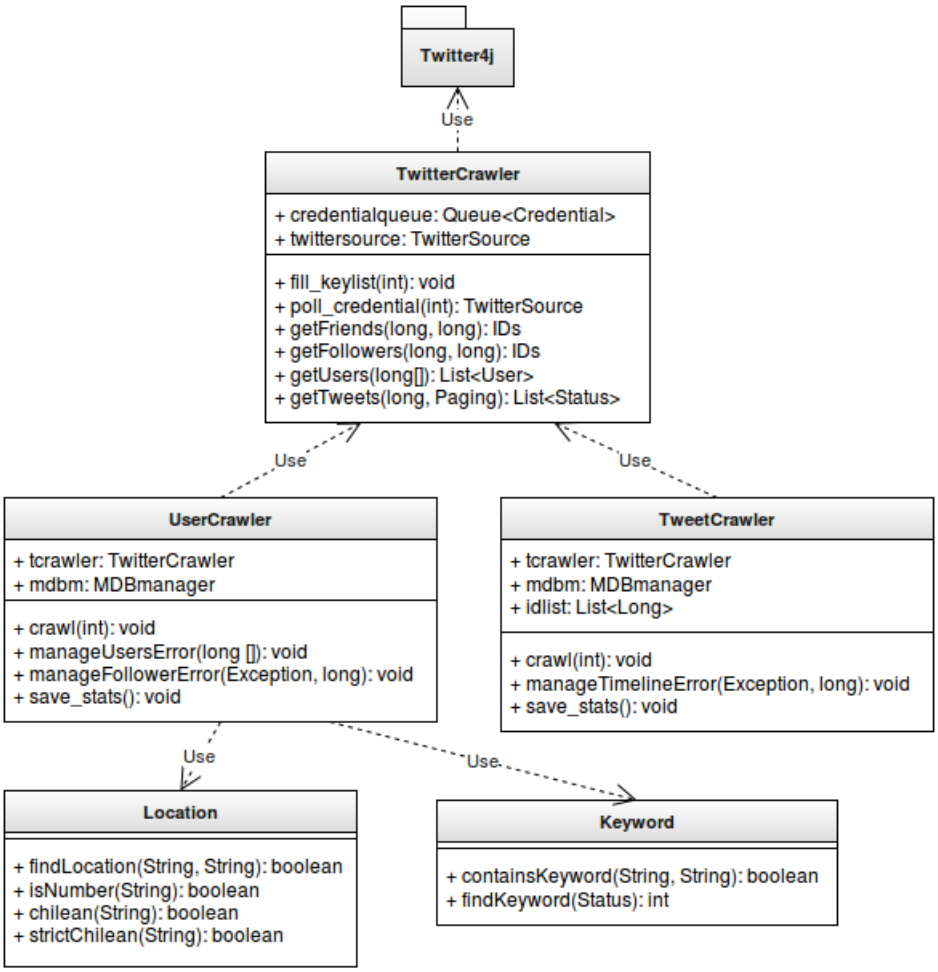

Figura 5.1: Diagrama UML de Clases del MCD
Fuente: Elaboración Propia

implementación de las ideas planteadas en el Capítulo 4, son puestas en práctica en el conjunto de clases mostradas en la Figura 5.1. A continuación son descritas en detalle.

- TwitterCrawler: los obstáculos impuestos por el uso de la *API Rest* de *Twitter* se resumen en cuatro puntos:
  1. Se necesitan credenciales para obtener los datos.
  2. El número de consultas tiene una cota máxima (por credencial).
  3. Este límite varía dependiendo del tipo de consulta.
  4. Las credenciales son reutilizables en periodos de 15 minutos.

  Esta clase es creada en respuesta a lo anterior, haciendo posible la obtención de datos de *Twitter* sin tener que preocuparse de las restricciones, ya que todo es gestionado automáticamente. De esta forma, son implementados elementos para el manejo de credenciales y la gestión de consultas. En efecto, son soportadas cinco peticiones:
  - Información de un usuario.
  - Listas de objetos con la información de los usuarios, en respuesta a un conjunto de IDs.
  - Lista de seguidores.

  - Lista de “amigos”.
  - El conjunto de *tweets* de cada usuario.

  Los elementos y métodos de esta clase son utilizados por las dos clases precisadas a continuación. Estas son los núcleos del Módulo de Recolección de Datos.

- UserCrawler: esta clase encarna la primera fase de recolección de información, es decir, el proceso de recorrido del grafo. Para hacer esto hace uso de colas y listas para mantener el orden los usuarios visitados y por visitar. Emplea una sección consignada para evaluar el criterio de nacionalidad y luego, otra para guardar la información del usuario, junto a los usuarios que lo siguen y a quienes él (o ella) sigue (“amigos”). La clase también maneja varios tipos de errores, ya que la recolección de datos depende de su disponibilidad, y varias excepciones originadas por la conexión de Internet y a los servidores de *Twitter*.
- TweetCrawler: aquí es implementada la segunda fase de recolección de datos, es decir, reunir los *tweets* que estén relacionados con marihuana de todos los usuarios almacenados en la fase anterior. Para hacer esto, es recorrida la lista de usuarios, luego es descargada y revisada su lista de *tweets*, para finalmente evaluar la presencia de alguna de las *keywords* asociadas con marihuana. Esta clase también implementa métodos de manejo de errores, similares a los de la clase anterior.
- Location: esta clase provee funciones para evaluar el criterio de nacionalidad. Se utiliza una cadena de caracteres como dato de entrada. Por lo tanto, se aplican métodos de tratamiento de texto y concordancia de palabras. Por otro lado, facilita dos métodos de evaluación de nacionalidad: uno estricto y otro suave. La diferencia radica en que el primero evalúa sólo la presencia de la palabra “chile”, mientras que el segundo determina la aparición del grupo enumerado en el Capítulo 4.
- Keyword: al igual que la anterior, esta clase proporciona métodos para evaluar la presencia de términos en el texto. En este caso, es usada la lista de 46 palabras relacionadas con marihuana, buscando en orden descendente de las mayormente utilizadas.

5.4.2. Módulo de Mantenimiento de Datos

En la práctica, el Módulo de Mantenimiento de Datos es el encargado de modelar la forma en que los datos están almacenados dentro del sistema y todas las interacciones con las bases de datos. No hay acción que se efectúe en las bases de datos sin que este módulo actúe como intermediario, a excepción del MVR, ya que el tiene su propia vía de comunicación.

En totalidad, la aplicación utiliza 4 bases de datos diferentes y sus respectivos modelos. Esta división se realiza debido a que están enfocadas a servir a etapas y módulos distintos del proyecto. Tres bases de datos están construidas en un gestor de bases de datos relacionales y la otra es una base de datos NoSQL. Esto se llevó a cabo por necesidades de rendimiento en el cálculo de algunas métricas de Análisis de Redes Sociales. Los modelos de las cuatro bases de datos son presentados en la Figuras 5.2, 5.3, 5.4 y 5.5. Las tres primeras son modelos Entidad-Relación que almacenan datos de distintas etapas:

- Minado de *Twitter*.
- Etiquetado de casos y entrenamiento de algoritmos.

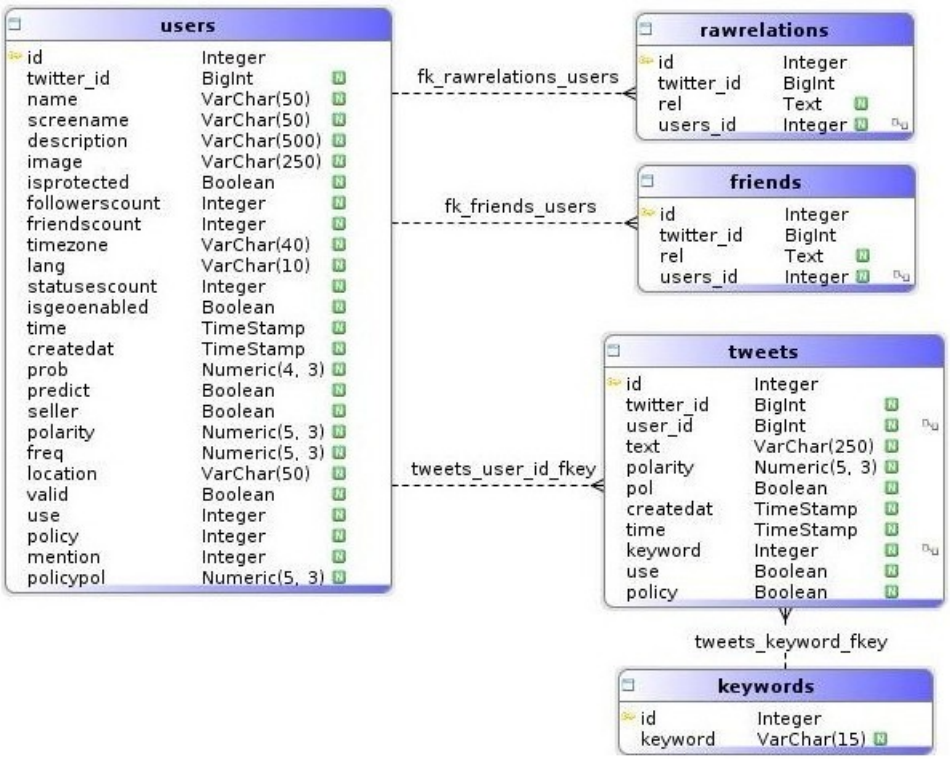

Figura 5.2: Modelo E-R de mdb
Fuente: Elaboración Propia

- Seguimiento a los indicadores a lo largo del tiempo.

La Figura 5.5 representa la base de datos NoSQL que funciona como una colección de documentos encargados de guardar las listas de seguidores y amigos en la red de usuarios “chilenos”.

El manejo de las bases de datos mediante código Java es posible a través de las correspondientes librerías de *PostgreSQL* y *ArangoDB*. Ambas permiten operaciones para crear, leer, actualizar y borrar elementos (CRUD, por sus siglas en inglés). Las clases MDBmanager, TDBmanager, DRmanager y TrDBmanager hacen uso de aquellas librerías y son las encargadas de prestar servicios a los módulos de la aplicación. Cada clase está encargada de gestionar una base de datos distinta. Principalmente, establecen conexiones a las bases de datos y realizan operaciones *CRUD*.

5.4.3. Módulo de Inteligencia de *Tweets*

El Módulo de Inteligencia de *Tweets* es empleado para adquirir información desde el texto consignado en los *tweets*. Esta tarea sostiene todas las métricas originadas desde la aplicación, por ende remarca la importancia de su implementación y en consecuencia, de su buen funcionamiento. El módulo arroja resultados con respecto a tres puntos de interés:

1. Polaridad del *tweet*.

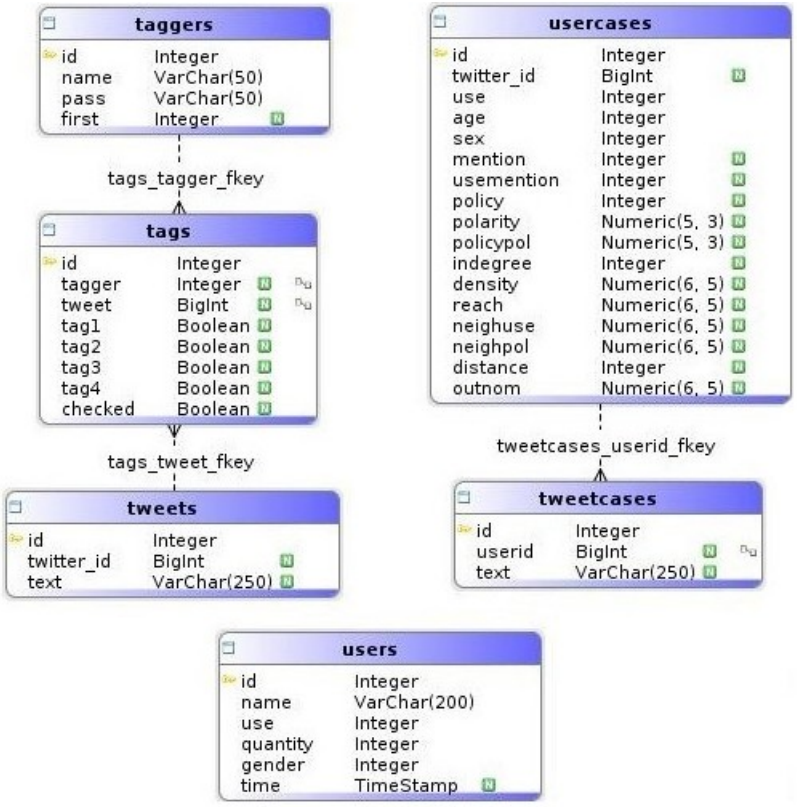

Figura 5.3: Modelo E-R de Tagging
Fuente: Elaboración Propia

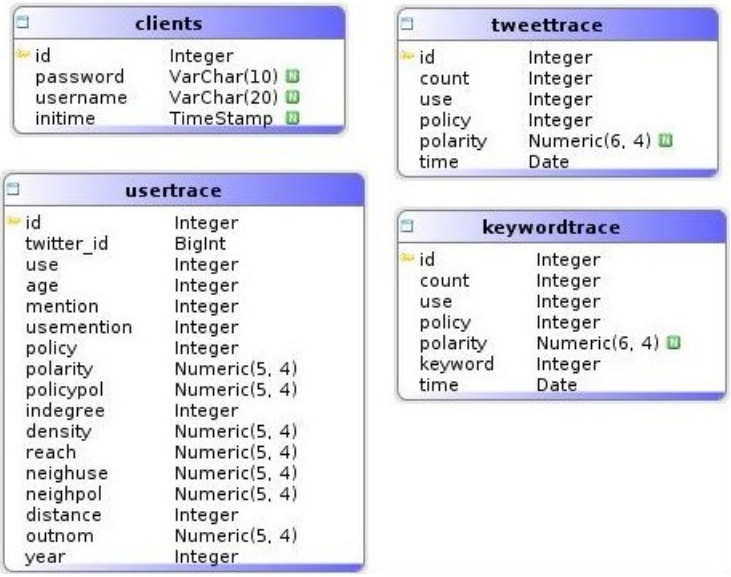

Figura 5.4: Modelo E-R de Trace
Fuente: Elaboración Propia

Figura 5.5: Modelo de Relations
Fuente: Elaboración Propia

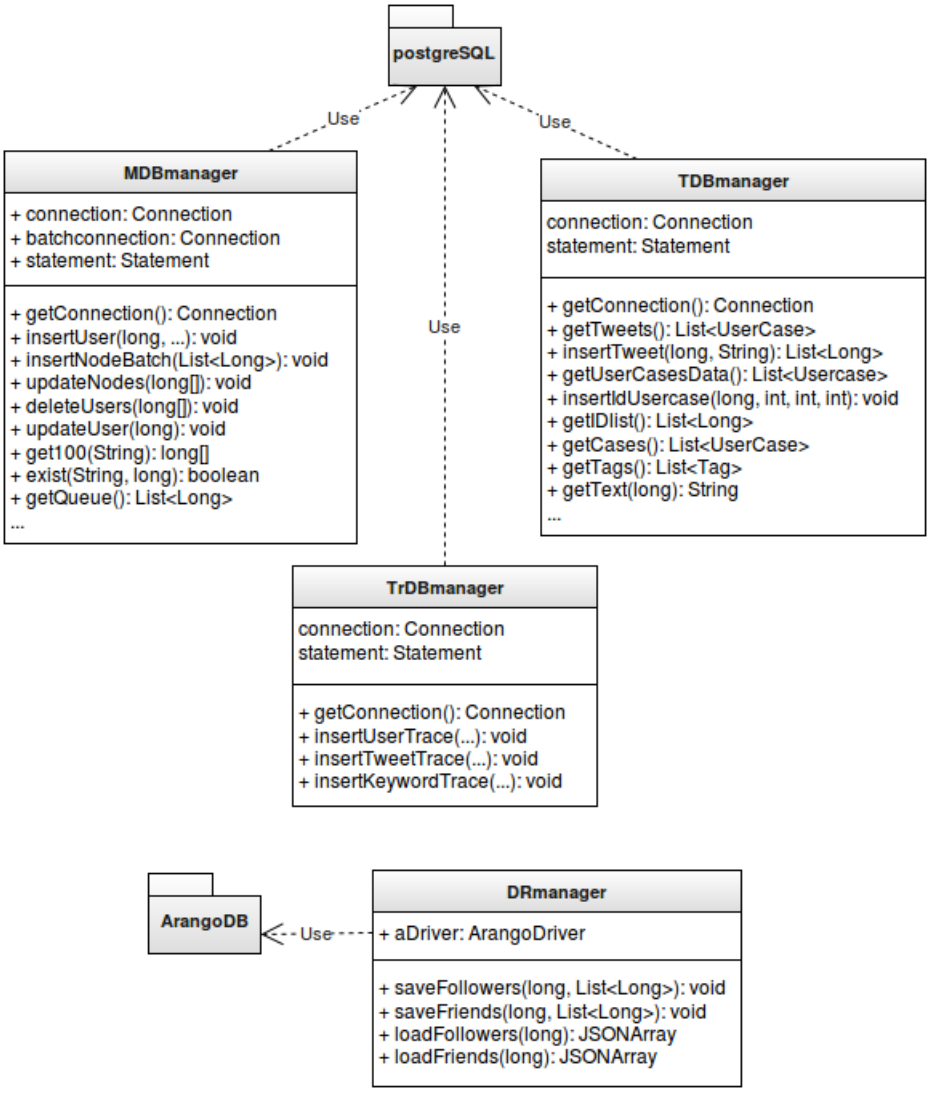

Figura 5.6: Diagrama UML de Clases del MMD
Fuente: Elaboración Propia

2. Consumo de marihuana por parte del autor.

3. Políticas públicas con respecto a la droga.

Trabajar con texto implica directamente que deben ser aplicadas herramientas que faciliten este proceso. Es aquí donde se hace uso de la librería de *Freeling*, específicamente su sección destinada a realizar *tokenización*, es decir, separación de los elementos del texto. El resto de servicios brindados por la librería no son aplicados a lo largo del proyecto. Por otro lado, se utiliza intensamente la librería de *Weka* para clasificar los textos dentro de las categorías binarias mencionadas anteriormente.

A continuación es detallado el funcionamiento de cada una de las clases representadas en la Figura 5.7, las cuales en unión conforman el MIT:

- Analyzer: esta clase se encuentra en representación de la librería de *Freeling*. Debido a que *Freeling* está escrito en lenguaje C++ no es posible llamar a sus métodos directamente, pero la misma librería brinda una API que funciona como intermediaria. Es así como la actual clase llama al único servicio puro de procesamiento de texto empleado en la aplicación, el servicio de *tokenización*.
- Preprocess: la clase Preprocess es la encargada de aplicar casi todos los métodos de preprocesamiento de texto. Los métodos también son utilizados por otros módulos que requieren procesamiento de texto, como por ejemplo el MIU que procesa los textos implicados en la predicción de edad. La clase Preprocess trabaja directamente con la clase anterior, realiza tratamiento a algunos elementos de *Twitter*, normaliza los textos, une cadenas de caracteres y transforma los textos en una sola línea. Todo esto es necesario para transformar la colección de *tweets* en una representación matricial.
- Polarity: aunque los procesos encargados de operar en los textos deberían ser aplicados conjuntamente, son separados por cuestión de velocidad. El cálculo de polaridad es realizado en una aplicación externa, denominada *PAPI*. Los resultados para cada texto son arrojados en intervalos de tiempo mayor a los clasificadores binarios, debido a que el análisis de la API es más complejo. La API es llamada mediante librerías URL de Java y por cada *tweet* se obtiene un objeto en formato JSON, conteniendo un valor real de polaridad. Finalmente, los resultados son guardados para su uso en las métricas.
- TweetClassifier: esta clase aplica los procedimientos estándar en el posicionamiento de los textos dentro de categorías. Estos son facilitados por la librería de *Weka*. El funcionamiento inicia con la carga de los modelos de clasificación y se establece la estructura de los datos a categorizar. Luego, cada *tweet* es confinado dentro de la estructura definida, clasificado con respecto al consumo y políticas de marihuana, y guardado dentro de la base de datos.

5.4.4. Módulo de Inteligencia de Usuarios

El Módulo de Inteligencia de Usuarios busca cumplir dos objetivos primordiales: calcular la prevalencia de marihuana a partir de datos obtenidos de *Twitter* y segmentar a los usuarios en base a su edad. El cumplimiento de esos dos puntos permitirá comparar los resultados del módulo con los datos producidos por la Encuesta Nacional de Drogas. En la práctica se verá que la búsqueda

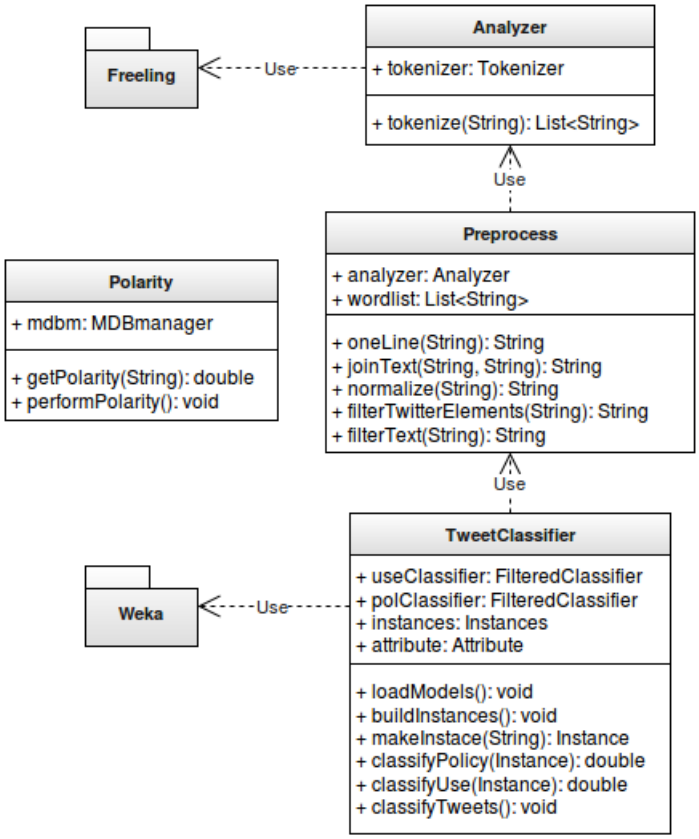

Figura 5.7: Diagrama UML de Clases del MIT
Fuente: Elaboración Propia

de esos objetivos brindará más cifras que ayudan a entender el consumo de marihuana.

Nuevamente será utilizada la librería de *Weka*, ya que permitirá cumplir con los dos objetivos, en otras palabras, permitirá clasificar el uso de marihuana y predecir la edad de un usuario de *Twitter*. Las dos tareas se cumplirán mediantes caminos diferentes y una servirá de alimento para otra. La predicción de edad conservará muchas de las propiedades aplicadas en el módulo anterior, pero la clasificación de consumo de marihuana sigue con los procedimientos clásicos de Minería de Datos.

La implementación del módulo está concentrado en tres clases, las cuales son representadas gráficamente en la Figura 5.8 y su relación. Las tres clases se involucran en la construcción de variables y la correspondiente aplicación de algoritmos de *Machine Learning*.

- Relations: esta clase se encarga de construir el arreglo de relaciones que serán utilizadas para el cálculo de métricas de Análisis de Redes Sociales. Primero son obtenidas las relaciones de seguimiento almacenadas por aplicación del Módulo de Recolección de Datos. Las relaciones guardadas están sujetas a interacción con toda la red de usuarios de *Twitter*, pero con motivo del cálculo de métricas egocéntricas, es necesaria la construcción de una red social acotada a la red propia de usuarios chilenos. Es así como en esta clase se obtiene la lista de usuarios seguidores y seguidos por cada usuario en la base de datos, y se cruza con toda esta para obtener la red social de usuarios chilenos. Finalmente, esta información es guardada en una colección de documentos brindada por *ArangoDB*.
- Egocentric: la clase Egocentric simboliza una etapa clave en el cálculo de prevalencia, porque es la que incorpora los estudios mencionados en el Capítulo 3. Aunque esta vez la red social es construida a partir de información de *Twitter*. Las variables son construidas desde la red personal de cada usuario, lo que le da el nombre a la clase. Se utilizan las listas de seguidores y "amigos" para construir siete variables:
  1. Densidad de vecindario.
  2. Grado de nominaciones entrantes.
  3. *Reach Centrality*
  4. Nominaciones externas
  5. Distancia mínima a autores de *tweets* de consumo.
  6. Porcentaje de autores de *tweets* de consumo en el vecindario.
  7. Polaridad promedio en el vecindario.

  La lista de seguidores es empleada para las tres primeras variables y la lista de "amigos" para el resto. Los algoritmos de cálculo de variables son de elaboración propia, porque las librerías de Análisis de Redes Sociales realizan los cálculos considerando la estructura total de usuarios. El tamaño de esta hace imposible computar las matrices de relaciones con recursos limitados. Además la clase permite el cálculo de métricas para años en particular. Esto se hace acotando el conjunto de *tweets* y usuarios considerados en el análisis.
- UserClassifier: la clasificación de edad y consumo de los usuarios se realiza en esta clase. Primero se cargan los modelos designados para realizar la predicción y se construyen las estructuras que permiten encapsular los casos. Debido a que el tiempo de procesamiento de las métricas egocéntricas es muy grande para considerar a todos los usuarios, se selecciona una muestra para cada año. Luego, para cada usuario son calculadas la métricas originadas en los *tweets* y en el Análisis de Redes Sociales. Por otro lado, son descargados los *tweets* del

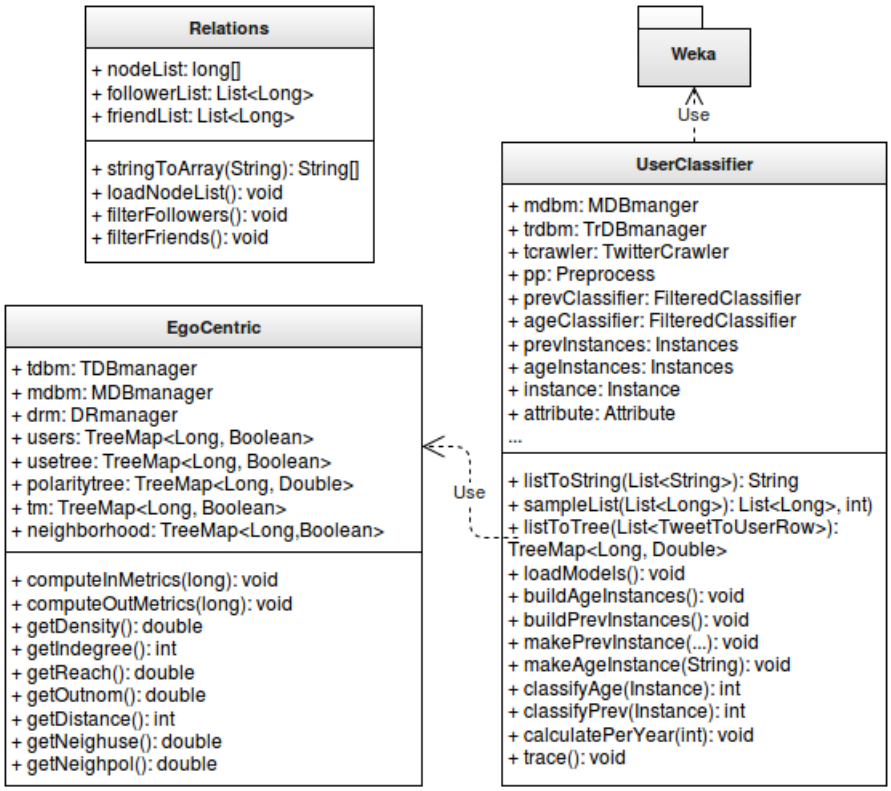

Figura 5.8: Diagrama UML de Clases del MIU
Fuente: Elaboración Propia

usuario para predecir la edad. Todas estas variables son encapsuladas dentro de una instancia para su correspondiente clasificación de consumo por parte del usuario.

5.4.5. Módulo de Evaluación de Rendimiento

El funcionamiento de este módulo es explicado mayoritariamente a su uso durante el desarrollo de la aplicación. Aunque puede ser empleado en elecciones futuras de algoritmos de clasificación, sólo es utilizado durante la elección del algoritmo más calificado en las tareas de clasificación y regresión. Los algoritmos elegidos son implementados en el Módulo de Inteligencia de *Tweets* y en el Módulo de Inteligencia de Usuarios.

En este apartado se utilizan los resultados del etiquetamiento de *tweets* y la encuesta contestada directamente por los usuarios de *Twitter*. El primero aporta los casos de estudio para el entrenamiento de clasificadores de texto y la segunda, ejemplos reales donde se observa el consumo de marihuana y la edad real de los usuarios.

El módulo es netamente dedicado a Minería de Datos y Minería de Textos, por lo tanto hace uso intensivo de la librería de *Weka*. En este apartado se logran vislumbrar muchas de las facilidades que ofrece la librería, abarcando casi todos las etapas del proceso KDD. Las clases que componen el módulo son mostradas en la Figura 5.9 y son detalladas a continuación:

- TweetFeatures: la clase se dedica a construir el conjunto de casos usados para el entrenamiento de algoritmos de clasificación de textos. Para hacer esto primero carga las etiquetas consignadas en la base de datos. Cada caso de *tweet* es asociado a su etiqueta respectiva, pre-procesado y agregado a un documento con formato que reconoce la librería *Weka*, ARFF.
- TextLearner: en el caso de texto ya se ha mencionado que requiere un tratamiento especial, consistente en transformar el conjunto de textos en una representación matricial. *Weka* utiliza una ligera modificación a esta representación, pero para casos prácticos se omitirá detalle. El tratamiento se aplica en forma de filtros, contando con varios parámetros para modificar, tal como el tipo de elemento de separación de textos, el número de *N-Grams*, normalización de la matriz, entre otros. Una vez obtenida la matriz se aplican filtros para desempeñar selección de atributos. Luego, se procede a evaluar el desempeño de las predicciones y regresiones con respecto a varios algoritmos. Las evaluaciones contienen las métricas convencionales de evaluación de algoritmos de aprendizaje. Finalmente, el modelo elegido se incorpora en un modelo y se guarda para su uso en otros módulos.
- UserData: esta clase tiene encomendada la tarea de constituir las variables que serán utilizadas en la clasificación de usuarios en el consumo de marihuana. Por lo tanto, comparte lógica con la clase UserClassifier del módulo anterior. Tan como su predecesor traspasa información desde los *tweets* a sus autores y computa las métricas egocéntricas. Luego procede a guardar los datos.
- UserFeatures: las variables necesarias para la clasificación y regresión (edad) son consignadas en esta clase. En primer lugar, son cargadas los casos de usuarios que serán utilizados en el entrenamiento. En segundo lugar, descarga los *tweet* de los usuarios etiquetados y realiza su correspondiente pre-procesamiento. Luego, guarda los casos de estudio. Por otro lado, también son cargadas todas las variables involucradas en la clasificación de consumo y son guardadas en el formato necesario para el entrenamiento.
- Learner: en el minado de datos para encontrar patrones que predigan el consumo de marihuana no es necesario un tratamiento especial, ya que a diferencia del texto, las variables ya están representadas de modo que la aplicación de algoritmos de entrenamiento es directa. Tampoco es necesaria la selección de atributos, ya que el tamaño no causa inconvenientes producidos por recursos limitados. Los datos son utilizados directamente en la evaluación de modelos. Son evaluados distintos algoritmos de clasificación y el elegido es guardado para su uso en el Módulo de Inteligencia de Usuarios.

5.4.6. Módulo de Visualización de Resultados

El Módulo de Visualización de Resultados fue implementado mediante una aplicación web, utilizando CodeIgniter como marco de desarrollo. No se involucra en la modificación de datos, ya que sólo está diseñado para mostrar los resultados interesantes desde el punto de vista del cliente. De esta manera, la función del módulo es ingresar a la base de datos en búsqueda de datos ingresados por la aplicación, transformarlos e implementar la lógica de visualización.

La aplicación web fue desarrollada bajo el enfoque promovido por CodeIngniter, el modelo MVC. Por lo tanto, la aplicación fue separada en en tres partes detalladas a continuación:

- Modelo: CodeIngniter facilita la conexión con diferentes sistemas de gestión de bases de

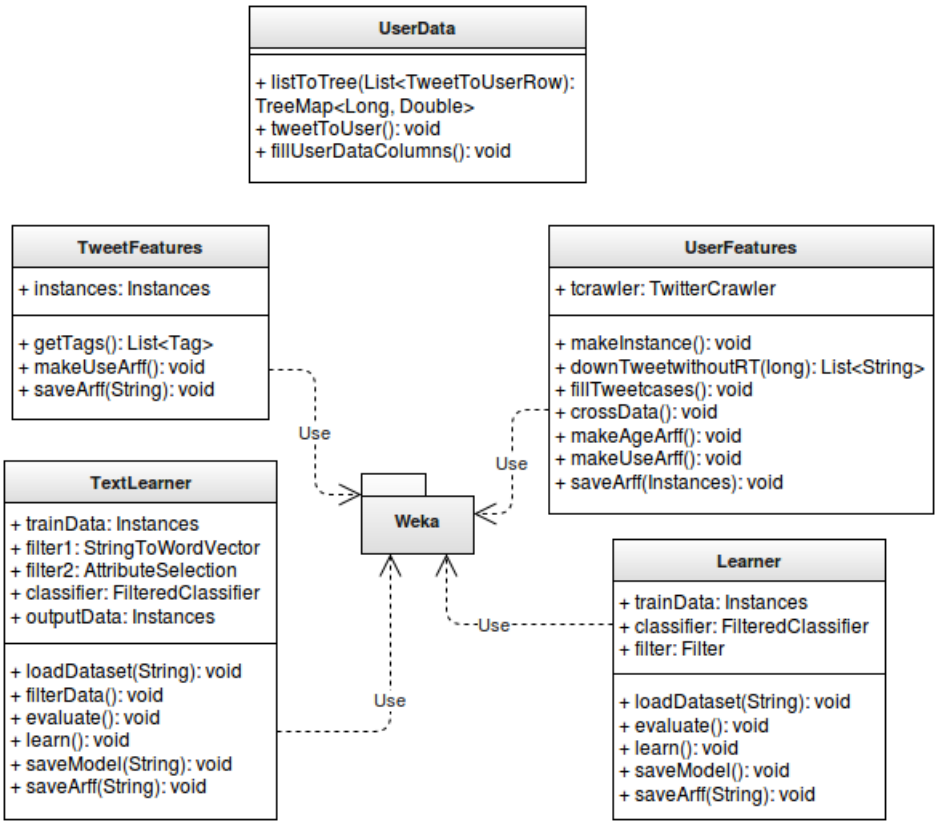

Figura 5.9: Diagrama UML de Clases del MER
Fuente: Elaboración Propia

datos, incluido PostgreSQL. En este caso se creó una conexión con la base de datos Trace, la que cuenta con la información del seguimiento de las métricas. Por lo tanto, se dotó de funciones para llamar a cada una de las métricas, separando por temporalidad. Por ejemplo, la polaridad puede ser calculada en forma diaria, mensual y anual. Los resultados son facilitados en forma de listas.

- Controlador: el controlador esta encargado de la lógica de negocio de la página web. Precisamente se implementó el manejo de sesiones para limitar el ingreso de la aplicación web de todo público. La figuras A.1 y A.2 del Anexo A muestran el proceso de validación. La tarea más importante del controlador es hacer manejo de los datos entregados por el modelo y procesarlos de forma correcta para su visualización. Para esto, administra las consultas requeridas a la base de datos, transforma los datos en arreglos y gestiona la navegación entre los datos a visualizar. Esto varía con respecto al tipo de métrica.
- Vistas: en la visualización abunda la presencia de conjuntos de datos. Por esta razón se utiliza *Chart.js*, una librería JavaScript para facilitar la visualización de gráficos de líneas y de barras. Los gráficos hacen uso de la herramienta de dibujo y brindan gráficos interactivos.

En el Anexo B se muestran algunas capturas de pantalla del prototipo implementado. La Figura B.1 muestra la vista de ingreso a la aplicación. En las figuras B.2, B.3, B.4, B.5 y B.6 se muestran respectivamente algunas vistas de las métricas. Cabe destacar que sólo la métrica de polaridad tiene implementada granularidad y la lógica de navegación por ella. Esta granularidad se divide en polaridad anual, mensual y diaria.

Capítulo 6

Resultados

En este capítulo será presentado el cúmulo de resultados que derivaron de la implementación. En su totalidad, son fruto de algunos módulos de la aplicación, así como de algunas actividades indispensables para el funcionamiento de la misma. Todos los datos brindan información relevante para comprender el fenómeno de *Twitter* y el consumo de marihuana dentro de ese contexto.

En primer lugar, se abordarán los resultados arrojados por la selección de palabras clave, que serán utilizadas en la recolección de *tweets*. A continuación de esto, se hará referencia a información originada en el Módulo de Recolección de Datos. Luego, serán mencionados algunos datos que fueron obtenidos en el etiquetado de *tweets* y usuarios. La evaluación de algoritmos no puede faltar, por lo que también tendrá destinada una sección del capítulo. Finalmente, se pondrá enfoque en las métricas elaboradas a partir de los resultados anteriores.

6.1. Palabras Clave

La lista original de palabras relacionadas con marihuana contenía un total de 80 palabras, todas ellas seleccionadas a partir de las fuentes mencionadas en el Capítulo 4. De ellas, 60 palabras fueron escogidas de la encuesta, en base a su frecuencia. Fueron procesadas por el modelo (*Topic Modeling*) sólo aquellas palabras que dificultaran la revisión manual, debido a su gran número de *tweets*. En la práctica sólo 29 palabras tenían más de 1000 *tweets*. Para cada palabra fue calculado el porcentaje de *tweets* efectivamente relacionados con marihuana y fueron conservadas aquellas que tuvieran una tasa mayor a cero. Algunas palabras continuaron en la ambigüedad, aún con la presencia de la cadena de caracteres “fum”, en cuyo caso fueron descartadas.

El proceso arrojó un total de 46 cadenas de caracteres, variando entre simples, bigramas y trigramas. Están divididos en dos grupos: cadenas únicamente relacionadas con marihuana y otras cadenas ambiguas, en las cuales se agrega el criterio mencionado anteriormente. La lista final de palabras clave se muestra en la Tabla 6.1.

Lista de Palabras Clave		
marihuana	cannabis	weed
mariguana	marijuana	prensada
porro (f)	thc	pito (f)
caño (f)	yerba (f)	sativa
sacate uno	canabis	macoña
de la buena (f)	hierba (f)	mota (f)
ganjah	cuete (f)	prensao
ganja	faso (f)	paraguaya (f)
de la wena (f)	cogollo (f)	bongazo
ganya	hachis	pitito (f)
matacola	hierva (f)	paragua (f)
marihuanita	troncho (f)	la verde (f)
canabica	cogollito (f)	pitits
cogoyo (f)	marimba (f)	paraguayo (f)
huiro (f)	bless (f)	yerva (f)
sacateuno		

Tabla 6.1: Palabras Clave
La etiqueta (f) indica la aplicación de la regla

6.2. Recolección de Datos de *Twitter*

La velocidad de recolección de *tweets* y usuarios del MCD es una medida importante en el funcionamiento de la aplicación, ya que condiciona la frecuencia en que los métricas pueden ser actualizadas. Ella depende de varios factores, tales como la velocidad de procesamiento de los recursos, la velocidad de Internet, el número total de nodos que fueron analizados, entre otros. Todo esto influye en la cantidad de usuarios que son evaluados y almacenados cada día.

La Figura 6.1 muestra el porcentaje de usuarios acumulados desde el día que inició la extracción. El total de usuarios al final del periodo de extracción fue de 1.505.367, aunque el número de usuarios válidos para el análisis fue de 1.361.285, debido al bloqueo de información por parte de ellos. Se pueden apreciar tres fases diferentes en la Figura 6.1, la primera corresponde a los cuatro primeros días, la segunda desde el quinto hasta el día 22 y la tercera desde el 23 hasta el final. La primera y la tercera son interesantes de analizar. La fase 1 tiene menor número de usuarios agregados al día porque desde el día cuatro fue implementada una mejora al algoritmo de recolección. En La fase 3 fueron encontrados pocos usuarios nuevos, esto marca el término natural del proceso.

La base de usuarios determina la información que puede ser extraída, debido a la fecha en que fueron consignados los datos. Por ejemplo, es imposible obtener métricas para fechas en donde no existían usuarios chilenos en *Twitter*. Lo anterior se ve reflejado en las Figuras 6.2 y 6.3, en donde la primera muestra el número acumulado de cuentas chilenas en *Twitter* para cada año y la segunda, revela el número de *tweets* creados para cada año.

La Figura 6.2 revela que para años anteriores al 2009 existían poco usuarios, por lo que desacredita resultados que puedan ser originados para esos años. La Figura 6.3 muestra la composición de

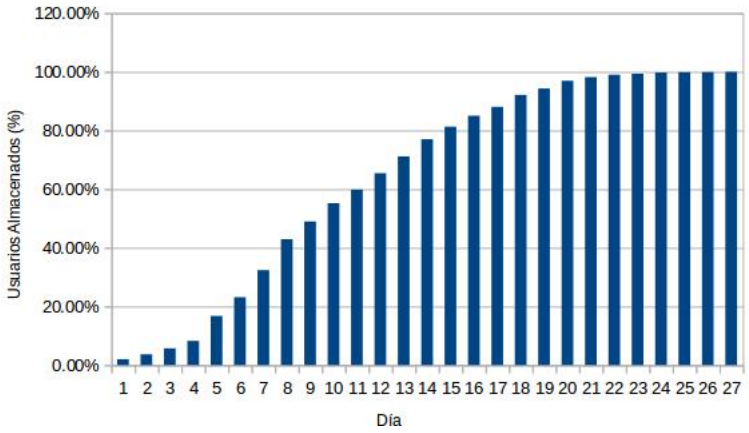

Figura 6.1: Gráfico de usuarios acumulados
Fuente: Elaboración Propia

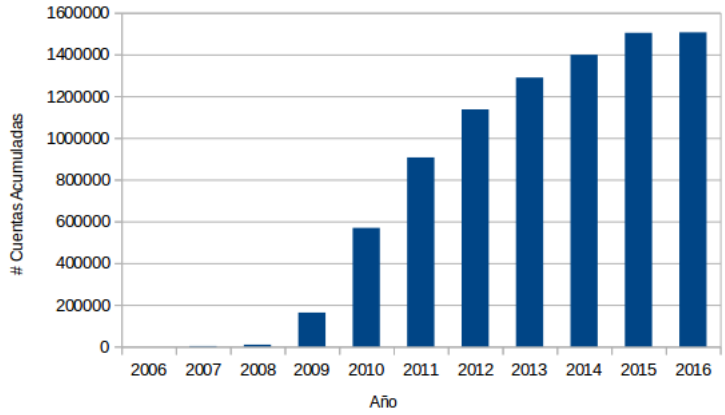

Figura 6.2: Gráfico de cuentas acumuladas
Fuente: Elaboración Propia

tweets en la base de datos, cuya forma está determinada por la restricción de *Twitter* de los último 3.200 *tweets* por usuario y el número de usuarios por año. En efecto, el grueso de los *tweets* se encuentra entre los años 2010 y 2016. Cabe recordar que sólo son almacenados los *tweets* relacionados con marihuana.

Los *tweets* no sólo son analizados desde el punto de vista de su distribución el tiempo, sino que también desde el potencial de generación por parte de los usuarios. En otras palabras, es interesante determinar cuantos usuarios están involucrados en la generación de la mayoría de los *tweets* ligados con marihuana. La curva de *Lorenz* de la Figura 6.4 explora esta idea, determinando que cerca del 10 % de los usuarios han producido el total de *tweets* almacenados en la base de datos, representando un total de 141.063 usuarios. Es aún más impresionante observar que cerca del 2 % de los usuarios generaron un 60 % de los datos, lo cual exhibe la desigualdad en la producción de textos de esta naturaleza. No se muestra gran parte del lado izquierdo del gráfico, ya que todos esos usuarios no tienen *tweets* asociados.

La aplicación también es nutrida por las relaciones existentes entre los usuarios de *Twitter*. Las relaciones de seguimiento son direccionales, es decir, sólo es necesario que algún usuario nomine

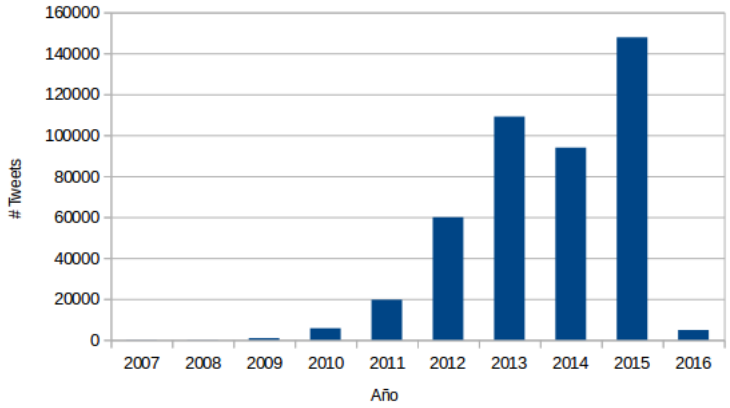

Figura 6.3: Número de *tweets* por año
Fuente: Elaboración Propia

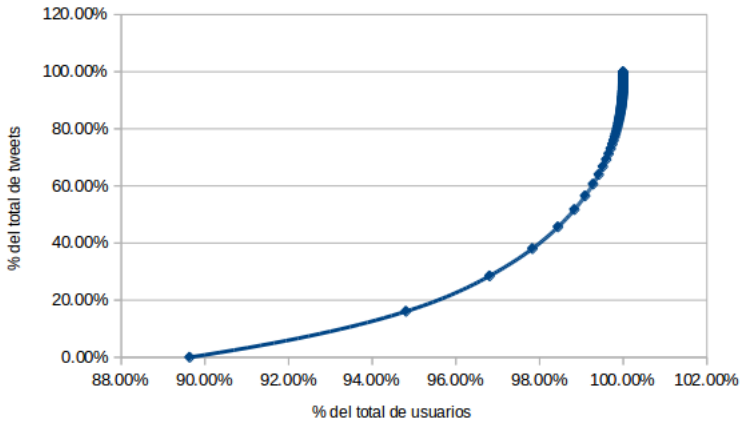

Figura 6.4: Curva de Lorenz de *Tweets*
Fuente: Elaboración Propia

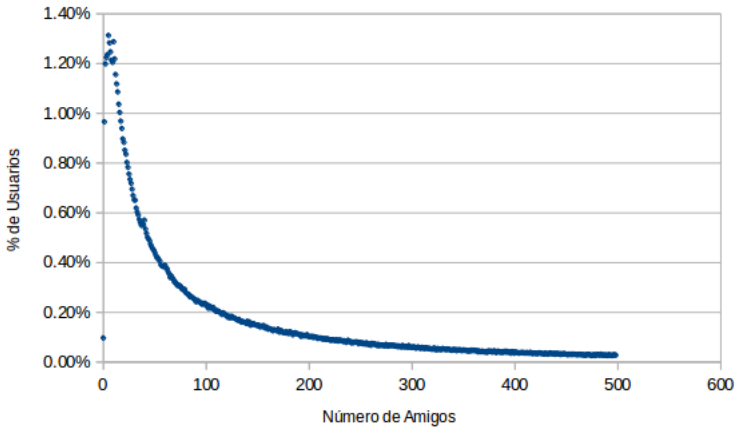

Figura 6.5: Distribución de Amigos
Fuente: Elaboración Propia

a otro para que la conexión exista. Esto convierte a la estructura en grafos direccionados y por consecuencia, pueden haber diferencias entre los seguidores y amigos. Hay que recordar que los amigos de un usuario son aquellos que él ha nominado.

La Figuras 6.5 y 6.6 muestran las distribuciones de amigos y seguidores, respectivamente. Estos figuras son gráficos de distribución de grado, *indegree* para los seguidores y *outdegree* para los amigos. En la Figura 6.5 observa que existen algunos usuarios con cero amigos y que gran parte de los usuarios siguen a menos de 500 cuentas, específicamente 87 % de los usuarios. En la Figura 6.6 se aprecia un comportamiento similar, la mayoría de los usuarios siguen a menos de 500 usuarios, pero en este caso es un 94 %.

La principal diferencia entre las distribuciones se sostiene en que la distribución de seguidores está más cercana a cero. Esto se debe a que las personas tienen la posibilidad de seguir a un gran número de cuentas, sin importar a quienes pertenezcan, pero el usuario promedio es seguido por un número más acotado. Probablemente personas conocidas fuera del del entorno de *Twitter*. Lo anterior apoya la forma en que opera el Módulo de Recolección de Datos, recorriendo la estructura de usuarios por medio de los seguidores. Ambas distribuciones están limitadas hasta 500 seguidores (o amigos), así es posible apreciar la forma de las curvas. Ambas se extienden a números más grandes, pero en esos límites las porcentajes comienzan a tender a cero.

6.3. Etiquetado

La clasificación y regresión de casos serán llevadas a cabo mediante algoritmos de aprendizaje supervisado, por lo tanto el etiquetado de casos condicionará totalmente cualquier resultado obtenido. Por esta razón, serán expuestos los principales resultados obtenidos de ambos procesos de etiquetado: el etiquetado de *tweets* y la encuesta a usuarios de *Twitter*.

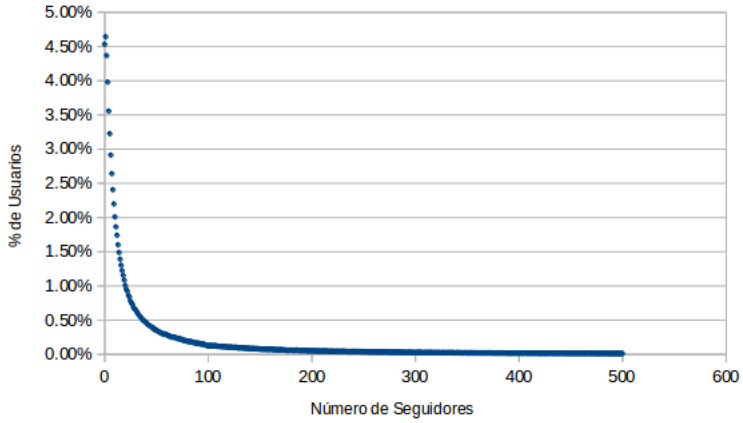

Figura 6.6: Distribución de Seguidores
Fuente: Elaboración Propia

Categoría	Acuerdo Relativo	Kappa de Fleiss
Ligado a marihuana	0.95	0.60
Consumo	0.89	0.56
Políticas	0.87	0.56
Venta	0.99	0.09

Tabla 6.2: Medidas de Acuerdo

6.3.1. Etiquetado de *Tweets*

A lo largo de los capítulos se han mencionado las categorías en que era necesario etiquetar los *tweets*. Tres de estas están dedicadas para el entrenamiento de algoritmos y otra para determinar la precisión de las palabras clave. El proceso de etiquetado de *tweets* arrojó un total 1.450 únicamente etiquetados y 50 etiquetados por cada una de las 12 personas. Es lógico empezar por los resultados obtenidos desde este último grupo, es decir, las medidas de acuerdo.

La Tabla 6.2 resume las medidas de acuerdo entre las 12 personas. Se observa un amplio nivel de acuerdo relativo para todas las categorías, todas superando el 0,95. Esta medida bruta es corregida para incorporar los efectos de la aleatoriedad en el proceso de etiquetado. Esto da como resultado el coeficiente *Kappa* de Fleiss, dedicado a reflejar el nivel de acuerdo entre más de dos personas. Tal como se visualiza en la Tabla 6.2, todas las categorías muestran un coeficiente cercano al 60 %, a excepción la categoría de venta. Esto se debe a que pesar de tener el nivel de acuerdo relativo más alto, los datos no tiene mayor variabilidad, por lo que el coeficiente es castigado directamente. Sin considerar esta categoría, las etiquetas cuentan con fuerza moderada de acuerdo.

Los 50 casos producen un dilema al momento de completar los 1.500 casos, ya que algunos reflejan contradicción entre las personas. Para solucionar esto se aproximó el promedio en cada uno de los 50. Los porcentajes para los casos positivamente clasificados se muestran en la Tabla 6.3. En ella se ve que el porcentaje realmente relacionado con marihuana es de 94,73 %, reflejando la precisión del procedimiento de búsqueda de palabras clave. Las categorías de consumo y políticas tienen heterogeneidad suficiente para el correcto entrenamiento de algoritmos. No así la categoría de venta, ya que sólo tiene heterogeneidad del 0,20 %, cantidad insuficiente para el entrenamiento

Categoría	Heterogeneidad
Ligado a marihuana	94,73 %
Consumo	13,80 %
Políticas	18,87 %
Venta	0,20 %

Tabla 6.3: Heterogeneidad en las etiquetas

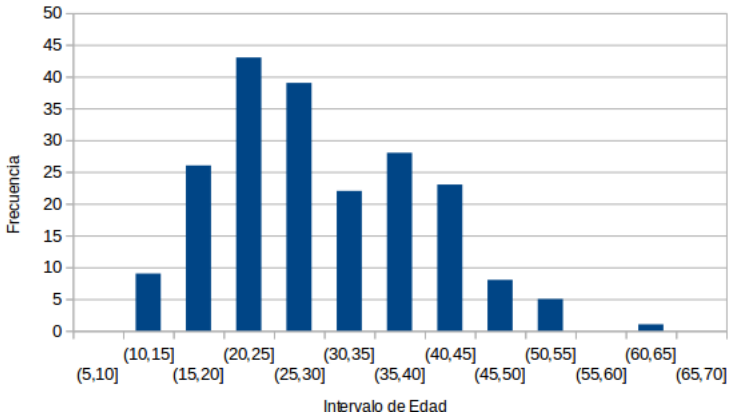

Figura 6.7: Distribución de edad de la muestra
Fuente: Elaboración Propia

de cualquier clasificador confiable. Por esta razón y su coeficiente *Kappa*, esta categoría es apartada de cualquier manipulación en etapas siguientes.

6.3.2. Encuesta a Usuarios

La encuesta a usuarios de *Twitter* fue realizada en el periodo comprendido entre el 9 de Febrero del año 2016 y el 6 de Marzo del mismo año. En ese periodo fue contestada por un total de 209 personas. Luego del cruce con la base de datos fueron obtenidos 204 casos factibles de uso, reflejando una tasa de respuesta del 0,3 %. Algunos casos descartados no fueron hallados en la base de datos y otros ni siquiera en una búsqueda manual por *Twitter*. Utilizando la tasa de consumo anual de marihuana (11,3 %) del año 2014 como tasa de heterogeneidad de los datos, se obtiene un error de 4,35 % y un nivel de confianza de 95 %.

Las 3 sencillas preguntas abordadas en la encuesta arrojaron algunas estadísticas para el análisis. En primer lugar, la edad de los casos recogidos brindan una visión corta de la distribución de edad de los usuarios chilenos de *Twitter*. La Figura 6.7 muestra dicha distribución, se aprecia claramente la ausencia de edades a los extremos. El intervalo con más presencia es entre los 20 y 30 años, y la edad media de la muestra es de 30,1 años. Por otro lado, 42,1 % de los usuarios reconocieron el consumo de marihuana en el último año y un 34,3 % en el último mes. Además un 43,6 % de los casos corresponden al sexo femenino. Lo antes mencionado es comparado con la edad media de 34,9 y el 51,4 % de mujeres obtenidos en el CENSO del año 2012.

Los datos obtenidos por la encuesta desprenden varios puntos interesantes para el estudio. El

primero va de acuerdo con la creencia de que los usuarios de *Twitter* son más jóvenes que la población general. El segundo sugiere la presencia de más hombres que mujeres, al contrario que los datos del CENSO. Cabe destacar que es muy probable que el tipo de encuesta esté presentando distorsiones. Sin ir más lejos es de esperarse que el alto porcentaje de prevalencia anual sea reflejo de la percepción que la encuesta estaba enfocada para consumidores de marihuana.

6.4. Evaluación de Algoritmos

En esta sección serán elegidos los algoritmos que cuenten con las mejores medidas de rendimiento para las tareas de clasificación y regresión. Los algoritmos elegidos serán incorporados como mecanismos centrales en los módulos de Inteligencia de *Tweets* e Inteligencia de Usuarios. En cada tarea serán evaluados varios algoritmos, seleccionando aquel que posea mayor poder predictivo.

Antes de presentar los números es necesario aclarar algunos puntos con respecto al procedimiento de evaluación y elección:

- En la práctica, la naturaleza de los datos condiciona directamente los algoritmos que pueden ser utilizados. En otras palabras, existe una diferenciación clara entre los algoritmos que pueden ser empleados en problemas de clasificación y en regresiones.
- La representación matricial de un grupo de textos produce una cantidad enorme de atributos. *Information Gain* fue la única técnica de reducción de atributos que fue empleada, ya que ha mostrado dar buenos resultados en texto. A pesar de esto, los algoritmos con mejor rendimiento no la incorporan, por lo que no será mencionada.
- Los algoritmos no son los únicos que varían en el proceso de prueba. Hay una serie de parámetros que pueden ser modificados, pero no será detallado su procedimiento de evaluación. El rendimiento de cada algoritmo incorpora intrínsecamente estas modificaciones, siendo sólo nombradas junto al algoritmo.
- Varios algoritmos de entrenamiento fueron descartados por su costos de empleo (tiempo de entrenamiento y exigencia computacional). Esto es aplicable a los textos, debido a la gran cantidad de atributos que generan.
- Si bien todas las medidas de rendimiento brindan información relevante acerca de la aplicación del algoritmo. En la elección se priorizará aquellos que posean mayor precisión para la clase de interés. Esto refleja la necesidad de recuperar casos en que efectivamente se evidencie el comportamiento.

6.4.1. Consumo en *tweets*

La evidencia de consumo de marihuana en *tweets* es abordado como un problema de categorización binaria, es decir, un problema de clasificación clásico. Por esto, existe una gran cantidad de algoritmos capaces de realizar la tarea. Aquí se evaluará la utilidad de tres algoritmos: *Naive Bayes* con monogramas y vectores de atributos binarios, *Voted Perceptron* con monogramas y vectores log-normalizados, y *Support Vector Machines* con monogramas a trigramas y vectores binarios.

Las Tablas 6.4, 6.5 y 6.6 muestran las medidas de rendimiento para cada uno de los algoritmos. En ellos se aprecia los valores de *Precision*, *Recall* y *F-Measure*. Los valores ponderados de todas las medidas se ven beneficiados de las altas cifras y la gran cantidad de casos para la clase cero. Como fue puntualizado anteriormente, se prioriza la *Precision* de la clase de consumo, por lo tanto es elegido el modelo de *Support Vector Machines*. El valor de *Recall* puede parecer poco, pero es compensado por su *Precision*, que si bien no es muy alto, es el mejor entre todos.

Clase	Precision	Recall	F-Measure
No consumo (0)	0,923	0,838	0,878
Consumo (1)	0,358	0,565	0,438
Ponderado	0,845	0,8	0,818

Tabla 6.4: Rendimiento de *Naive Bayes* para el consumo en *tweets*

Clase	Precision	Recall	F-Measure
No consumo (0)	0,88	0,971	0,923
Consumo (1)	0,486	0,174	0,256
Ponderado	0,826	0,861	0,831

Tabla 6.5: Rendimiento de *Voted Perceptron* para el consumo en *tweets*

Clase	Precision	Recall	F-Measure
No consumo (0)	0,883	0,978	0,928
Consumo (1)	0,588	0,193	0,291
Ponderado	0,843	0,87	0,84

Tabla 6.6: Rendimiento de SVM para el consumo en *tweets*

6.4.2. Políticas en *tweets*

Al igual que en la parte anterior, la presencia de políticas relacionadas con marihuana en los *tweets* es un problema clásico de clasificación binaria. A pesar de ello, en esta oportunidad el conjunto de algoritmos es diferente: SVM con monogramas a trigramas y vector log-normalizado, *Voted Perceptron* con monogramas y vector log-normalizado, y Árbol de Decisión C4.5 con monogramas y vector binario.

Las Tablas 6.7, 6.8 y 6.9 muestran las cuatro medidas de rendimiento para cada uno de los algoritmos bajo mira. Las tres alternativas tienen valores parecidos en las medidas de la clase cero y la ponderación para ambas clases. Por ende, el factor diferenciador está en las métricas de la clase 1. Nuevamente la decisión reside en la mayor *Precision*, es decir, *Voted Perceptron*. Cabe destacar que los otros dos algoritmos tienen asociados mayores valores de *Recall*, pero son menospreciadas a cambio del valor antes mencionado.

El mejor modelo para la clasificación de políticas resulta ser considerablemente mejor que su par de consumo. De hecho, es 0,23 veces mejor en *Precision*. Esto implica que la presencia de elementos que permitan determinar si un *tweet* corresponde a políticas relacionadas con marihuana es más clara. Pudiendo ser necesario más contexto para determinar de manera certera si un *tweet* menciona consumo de marihuana. Cabe recordar que se está tratando de categorizar textos que ya están relacionados con la droga.

Clase	Precision	Recall	F-Measure
No políticas (0)	0,865	0,967	0,913
Políticas (1)	0,714	0,353	0,473
Ponderado	0,837	0,851	0,83

Tabla 6.7: Rendimiento de SVM
para políticas en *tweets*

Clase	Precision	Recall	F-Measure
No políticas (0)	0,86	0,971	0,912
Políticas (1)	0,722	0,322	0,445
Ponderado	0,834	0,849	0,824

Tabla 6.8: Rendimiento de *Voted Perceptron*
para políticas en *tweets*

Clase	Precision	Recall	F-Measure
No políticas (0)	0,874	0,946	0,908
Políticas (1)	0,639	0,413	0,502
Ponderado	0,83	0,845	0,832

Tabla 6.9: Rendimiento de C4.5
para políticas en *tweets*

6.4.3. Edad de Usuarios

La predicción de edad comparte la misma base de las clasificaciones anteriores. En el sentido de que utiliza elementos del lenguaje para reconocer parámetros ocultos que determinen la edad de las personas. Se apoya en la percepción de que los individuos cambian el conjunto de palabras que ocupan a lo largo de su vida y que generaciones enteras comparten elementos léxicos.

El elemento novedoso de esta parte radica en que ya no se trata de encapsular los textos dentro de categorías, sino que se intenta asociar a los casos dentro de un rango de valores. Esta diferencia también se ve reflejada en las métricas de rendimiento que se ocuparán. En esta ocasión serán utilizadas medidas de relación lineal y diferencias agregadas entre los datos reales y los predichos. Específicamente se utilizará la correlación de *Pearson* y otros errores.

En esta instancia son evaluados tres algoritmos diseñados para realizar regresiones de datos: Regresión Lineal, M5P y la versión de *Support Vector Machines* para regresiones. Todas fueron

Modelo	Correlación	MAE	RMSE
Regresión Lineal	0,248	7,913	9,792
M5P	0,469	7,286	9,234
SVMreg log-normalizado	0,526	6,573	8,503
SVMreg binario	0,583	6,280	8,151

Tabla 6.10: Rendimiento de algoritmos de edad

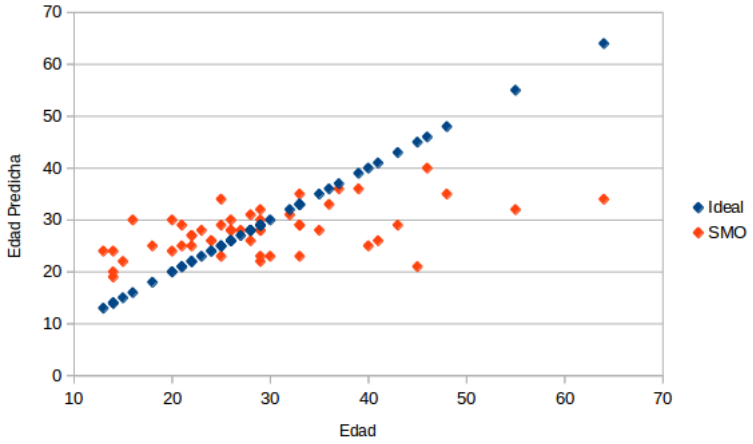

Figura 6.8: Evaluación gráfica de SVM
Fuente: Elaboración Propia

entrenados con monogramas y vectores de frecuencias log-normalizados. Aunque el mejor modelo tiene una pequeña variación. Las medidas de rendimiento se muestran en la Tabla 6.10, ahí figuran todas las opciones más una versión de SVM con vectores binarios. Se puede apreciar claramente que las medidas mejoran estrictamente de arriba hacia abajo. El mejor modelo es la última versión de SVM, teniendo una correlación de *Pearson* de 0,583 y error absoluto medio de 6,28. En otras palabras, el modelo se equivoca en en promedio cerca de 6 años.

La Figura 6.8 muestra gráficamente los resultados del modelo para un conjunto de prueba del 25 %, comparando lo resultados del modelo con respecto a la recta identidad. En ella se puede ver que a medida que avanza la edad el error incrementa, pudiendo deberse a que el léxico de las personas se estabiliza desde cierta edad. Por otro lado, se debe destacar la mejora del modelo al considerar la presencia de palabras en el vocabulario en vez de la frecuencia. Esto indica que la edad de una persona está mayormente determinada por el conocimiento y uso de ciertas palabras.

6.4.4. Consumo de Usuarios

La clasificación de consumo de marihuana por parte de los usuarios se incluye junto a los típicos modelos de Minería de Datos, debido que aquí no se hará tratamiento de textos para conseguir un conjunto de atributos. El grupo de 13 atributos, compuesto por medidas derivadas de los *tweets* y el entorno social del usuario, será utilizado para predecir su consumo de marihuana.

Fueron utilizados tres algoritmos para comparar sus rendimiento en la clasificación: *Support*

Clase	Precision	Recall	F-Measure
No consumo (0)	0,757	0,709	0,732
Consumo (1)	0,628	0,684	0,665
Ponderado	0,703	0,698	0,7

Tabla 6.11: Rendimiento de SVM
para consumo en usuarios

Vector Machines, *Multilayer Perceptron* y *Voted Perceptron*. Los tres en sus versiones optimizadas arrojaron medidas casi idénticas, sólo variando en la medida *ROC Area*. Sugiriendo que los tres algoritmos exprimen casi todo el poder predictivo del conjunto de variables. La Tabla 6.11 muestra el conjunto de medidas de rendimiento para SVM, utilizado por defecto como modelo final, porque permite apreciar la influencia de las variables en la clase. Estos valores se diferencian ampliamente a modelos anteriores, ya que las medidas están balanceadas. Esto se cumple para las dos clases y para los valores de *Precision* y *Recall*. En síntesis, individualmente será recuperado el 68,4 % de los consumidores de marihuana y 62,8 % del total de predichos serán efectivamente consumidores en el último año.

La Tabla 6.12 muestra los pesos normalizados para cada variable. Los datos indican que las variables con mayor poder predictivo son la edad, la emisión de *tweets* relacionados con marihuana, *tweets* sobre políticas de marihuana, el porcentaje de consumidores en el vecindario personal y las nominaciones fuera de la red social. Específicamente, la primera y la quinta disminuyen el riesgo de consumo, y la segunda, la tercera y la cuarta lo aumentan. En contraposición, la polaridad y la densidad son las variables más débiles.

Atributo	Peso Normalizado
Edad	−1,58
Tweets de marihuana	2,67
Consumo en *tweets*	0,51
Políticas en *tweets*	1,50
Polaridad	−0,14
Polaridad de Políticas	0,68
Seguidores	0,27
Densidad	0,05
Reach Centrality	0,95
Uso en vecindario	1,37
Polaridad en Vecindario	0,31
Distancia a consumidores	0,30
Nominaciones externas	−1,80
Intercepto	0,13

Tabla 6.12: Influencia de variables en el
consumo de marihuana

Se confirman varias creencias y se replican algunos resultados obtenidos en la literatura. En primer lugar, los consumidores se concentran en segmentos de edad más jóvenes. Las popularidad de una persona aumentan su riesgo a consumir marihuana. El comportamiento del entorno influencia directamente al comportamiento de las personas. Bajo este contexto, la emisión de *tweets* de

consumo de los amigos predice con mayor fuerza que la emisión propia. La publicación de cualquier mensaje relacionado con marihuana dice mucho del consumo. Además, todas las medidas de cercanía a otros consumidores aumentan el riesgo de consumo.

6.5. Métricas

En esta sección serán mostrado le producto final de esta memoria. Estos resultados están obtenidos a partir de todo lo realizado anteriormente, y ayudarán a comprender el fenómeno de la marihuana en *Twitter*. Incluso permitirán sacar conclusiones de su relevancia para la población general. Los resultados son desplegados en orden distinto que el contenido en el Capítulo 3. En esta ocasión se procederá de la siguiente manera: prevalencia, frecuencia de consumo, polaridad, polaridad de políticas, porcentaje de consumidores entre amigos, oferta de marihuana, y palabras utilizadas en *tweets* de consumo.

6.5.1. Prevalencia

En epidemiología, la prevalencia representa el porcentaje de la población que evidencia cierta característica en un periodo de tiempo. En este caso se trata de perseguir la misma definición, pero desde datos consignados en *Twitter*. Para obtener esta métrica es necesaria toda la estructura de la aplicación, desde los recolectores de información de *Twitter* hasta el clasificador de consumo. Este último es utilizado para determinar el consumo de marihuana en el último año para una muestra de usuarios.

La Figura 6.9 muestra el cálculo de prevalencia para cada año, la cual revela un alto porcentaje para el años anteriores al 2010. Los valores para esos años pueden estar sobrestimados debido a los pocos datos de usuarios y *tweets* para ese periodo, y la utilización un supuesto clave: las relaciones entre usuarios creados en ese tiempo no han cambiado drásticamente al avanzar los años. Los años posteriores al 2009 muestran una evolución paulatina del consumo de marihuana entre los usuarios chilenos de *Twitter*.

Un análisis necesario, para determinar la representatividad de los datos con respecto a la población chilena, es realizar una comparación entre los datos mostrados en la Figura 6.9 y los datos recogidos por la Encuesta Nacional de Drogas. La prevalencia histórica arrojada por esta encuesta es expuesta en la Figura 6.10. Se puede apreciar una similitud entre la tendencia de los años 2010 y 2016 de la curva producida por el predictor y la tendencia entre los años 2008 y 2014 del estudio nacional, aunque se tiene un desfase de dos años. La Figura 6.11 grafica la comparación de curvas para el periodo entre 2008 y 2014, corrigiendo a la curva predicha por un ponderador y desplazándola dos años atrás para corregir el desfase.

La gran similitud entre curvas es innegable, presentando un coeficiente de correlación de *Pearson* de 0,933. Aunque hay que destacar que la curva predicha fue retrasada en dos años. Este desfase podría estar producido por las variables utilizadas en el predictor, es decir, el resultado está totalmente condicionado a elementos presentes en *Twitter*. Esto quiere decir que el consumo

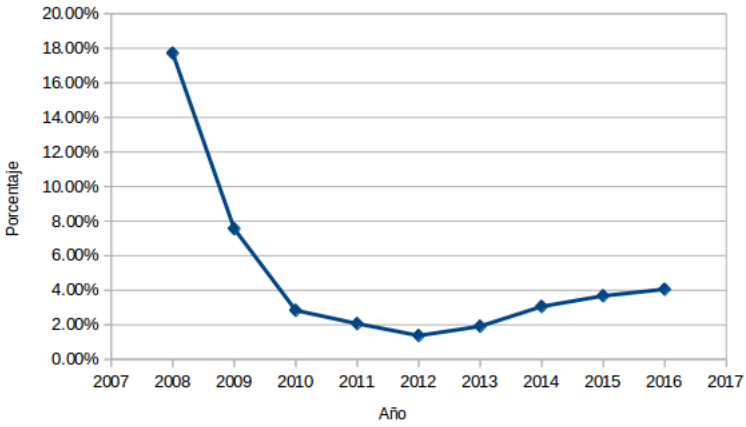

Figura 6.9: Prevalencia Anual
Fuente: Elaboración Propia

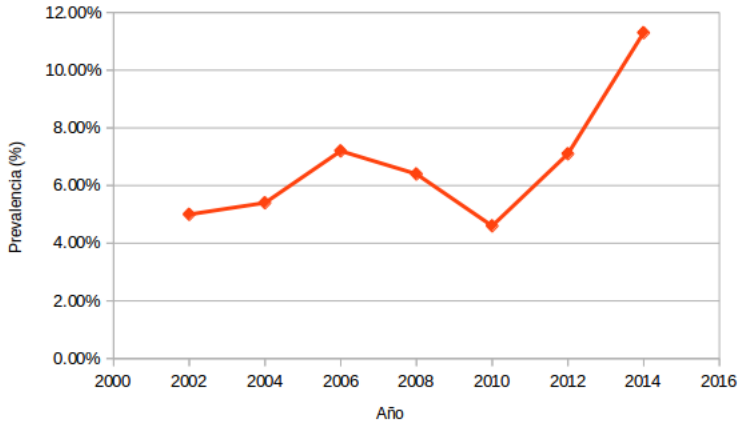

Figura 6.10: Prevalencia Nacional
Fuente: Estudio Nacional de Drogas

de marihuana no es reflejado inmediatamente en el contexto de *Twitter*, ya que requiere que los usuarios presenten el comportamiento y luego generen contenido que esté relacionado con él. El desfase será replicado para el análisis de otras métricas.

6.5.2. Frecuencia de Consumo

La capacidad de reconocer si un *texto* contiene elementos que permitan predecir el consumo de marihuana es uno de los ejes centrales de la aplicación. Puede aportar en el cálculo de prevalencia, pero también produce una métrica por sí mismo, el cálculo de frecuencia de consumo. Este se construye promediando los *tweets* de consumo por año de quienes emiten este tipo de textos. La Figura 6.12 reproduce el cálculo de esta métrica para los años pertinentes. En ella se puede apreciar un alza sostenida.

Fue realizado el mismo procedimiento de desfase y escalamiento visto en la métrica de prevalencia. Esta medida replicó de alguna manera los resultados obtenidos anteriormente, ya que produjo

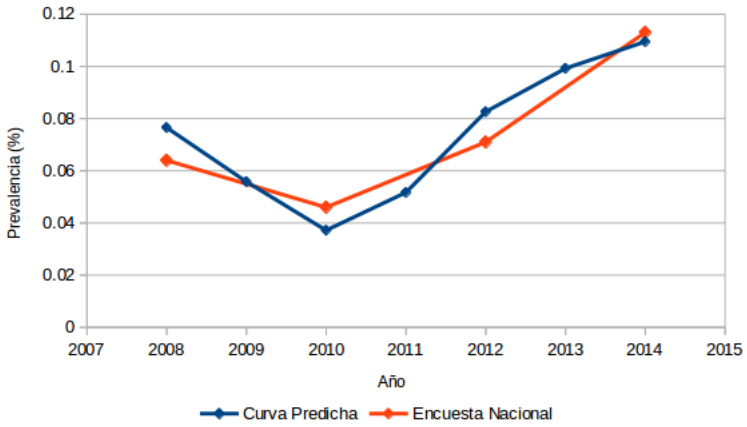

Figura 6.11: Comparación de Prevalencia
Fuente: Elaboración Propia

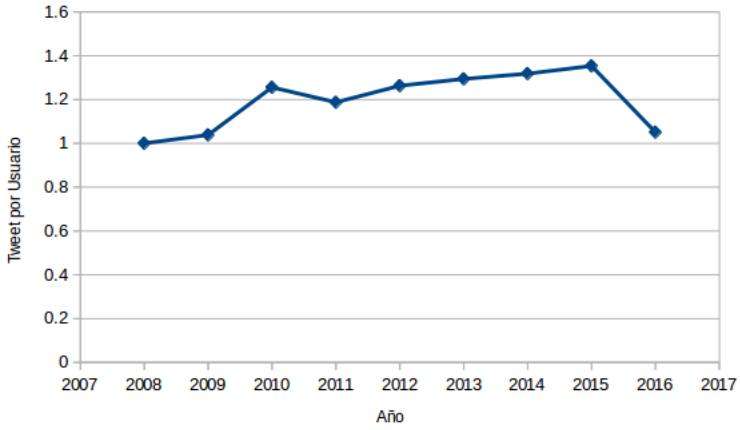

Figura 6.12: *Tweets* de Consumo por Usuario
Fuente: Elaboración Propia

curvas sumamente parecidas. En la Figura 6.12 se grafican las dos curvas llevadas a la escala de días de consumo, ya que la Encuesta Nacional de Drogas mide el número de días en el último mes que se consumió la droga. En este caso, el coeficiente de correlación de *Pearson* es igual a 0,471. De esta manera, esta métrica también presenta mejoras en la correlación a partir de un desfase de dos años.

6.5.3. Polaridad

La polaridad de *tweets* refleja que tan positivos o negativos son los textos emitidos por los usuarios de *Twitter*. Esta métrica trata de incorporar las opiniones vertidas en el *tweet* promedio para cada periodo y así, realizar seguimiento al efecto en la opinión de las personas a partir de ciertos eventos. La polaridad es calculada para cada año, mes y día para los cuales se poseen datos. La Figura 6.14 exhibe la evolución anual para esta métrica. En una primera instancia sólo se mencionará su forma, evidenciando una baja desde el año 2008 y recuperándose desde el año 2013.

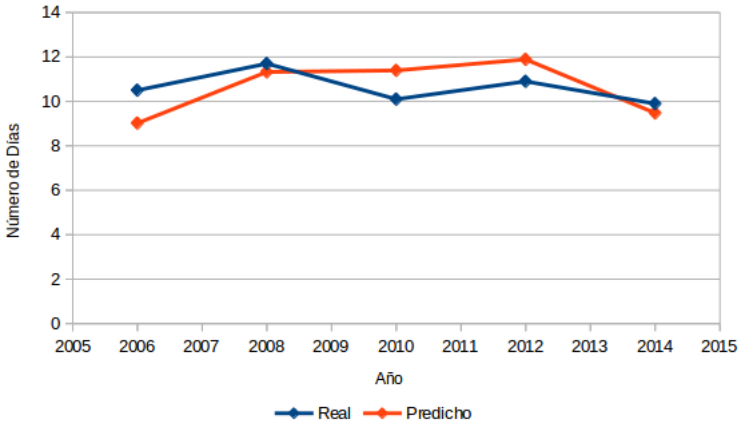

Figura 6.13: Frecuencia de Consumo
Fuente: Elaboración Propia

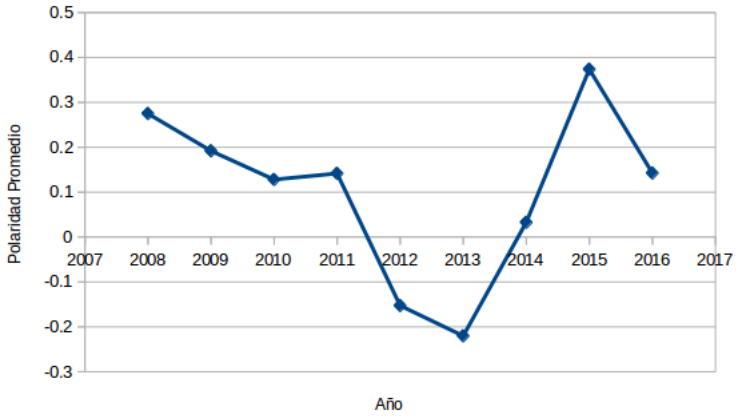

Figura 6.14: Polaridad de *Tweets*
Fuente: Elaboración Propia

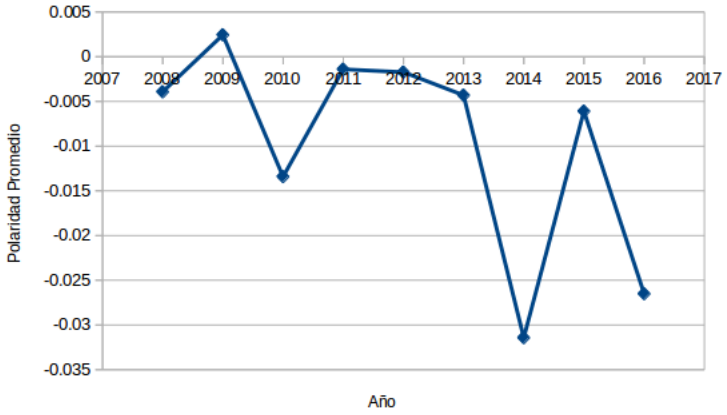

Figura 6.15: Polaridad de Usuarios
Fuente: Elaboración Propia

La polaridad también es calculada en base a los usuarios. Cada usuario tiene un número de *tweets* asociados y ellos, una polaridad. Se calcula tomando el promedio entre los *tweets* del usuario y luego, el promedio de los usuarios. Esto implica que gran cantidad de usuarios tendrán polaridad igual a cero. La Figura 6.15 muestra la evolución de la polaridad de usuarios a través de los años, detectándose una baja sostenida. Es importante destacar la diferencia de forma entre los dos gráficos de polaridad, señalando que la evolución diaria de los *tweets* no es incorporada directamente en la polaridad de usuarios.

En un momento se planteó que la polaridad de marihuana podría estar relacionada con la percepción de riesgo de la droga. Obedeciendo la definición de percepción de riesgo corresponde al porcentaje de la población que considera riesgoso el consumo experimental o frecuente de marihuana. Para efectos del análisis se considerará sólo el segundo. Por otro lado, es lógico pensar que esta medida tiene similitud con el promedio de polaridad para *tweets* negativos. En otras palabras, las mismas personas que opinan negativamente de la droga también la consideran riesgosa. La Figura 6.16 la explora de esta idea, comparando la percepción de riesgo con el promedio de polaridad negativa de *tweets*. Esta última transformada mediante una reflexión con respecto al eje horizontal, retrasada en dos años y escalada. El coeficiente de correlación de *Pearson* es de 0,819, evidenciando un gran parecido tanto gráfico como numérico y apoyando nuevamente la teoría del desfase.

6.5.4. Polaridad de Políticas

La polaridad de *tweets* de políticas relacionadas con marihuana comparte el mismo principio que su par mencionado anteriormente, pero esta vez es aplicado sólo a *tweet* clasificados como políticas. La Figura 6.17 muestra la curva de esta métricas a lo largo de los años. Se hayan diferencias claras con respecto al gráfico de polaridad general de *tweets* de marihuana. Por otro lado, es inevitable notar la relación entre el aumento de polaridad de los últimos años y el toda la atención mediática que ha sufrido la marihuana en casi el mismo periodo. También es importante notar la similitud de la curva con la apreciada para la prevalencia. Esta relación es apoyada por el modelo predictor de consumo en usuarios.

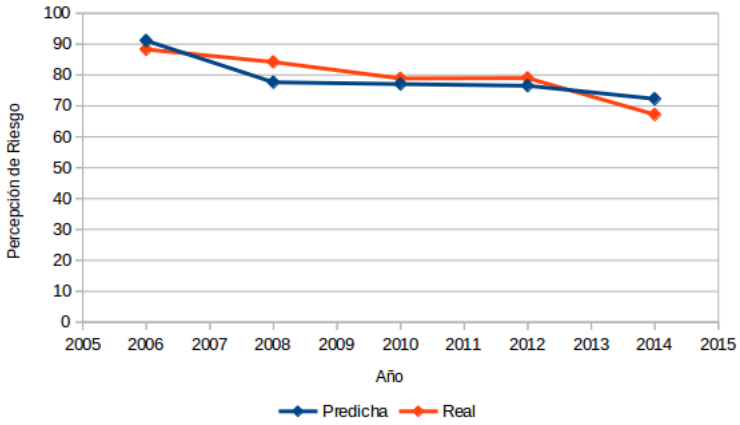

Figura 6.16: Comparación de Percep. de Riesgo
Fuente: Elaboración Propia

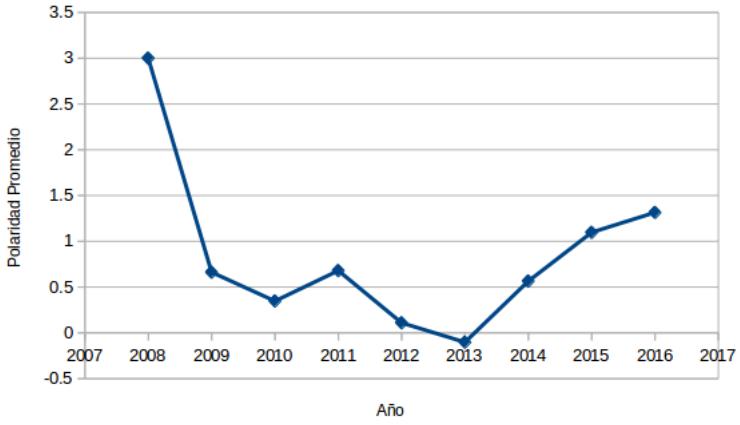

Figura 6.17: Polaridad en *Tweets* de Políticas
Fuente: Elaboración Propia

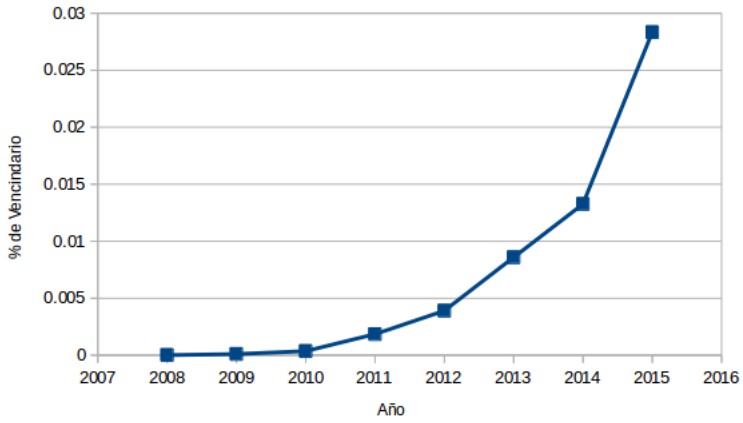

Figura 6.18: Vecindario Consumidor
Fuente: Elaboración Propia

6.5.5. Amigos Consumidores

El porcentaje de amigos presentes en la red egocéntrica del usuario es más un atributo que una métrica en sí. Su valor radica en el poder predictivo del consumo de marihuana. Aún así, como fue planteado en el Capítulo 3, podría brindar nociones de facilidad de acceso. Tal como se puede apreciar en la Figura 6.18, evidencia un incremento drástico desde que existen datos. Para valorar esta métrica, se debe tener en cuenta el supuesto detrás de las relaciones de seguimiento para el transcurso de los año.

6.5.6. Oferta de Marihuana

La métrica de oferta de marihuana se basa en la clasificación de los *tweets* con respecto a venta de marihuana. Como se vio en la práctica, los *tweets* etiquetados como venta representaban una porción muy pequeña como para ser tomada en cuenta. Esto da como resultados un coeficiente de *Kappa* muy bajo y no permite el correcto entrenamiento de algoritmos. Durante la revisión manual de algunos *tweets* se reconoció claramente la venta por parte de usuarios, pero la muestra de 1.500 textos para el etiquetado no los incorporó en forma suficiente. Esto indica que existe, pero es necesario revisarlos con mayor profundidad para reconocerlos automáticamente. La métrica de oferta de marihuana fue la única que no pudo ser implementada.

6.5.7. Palabras Utilizadas

Finalmente, es turno de reconocer las palabras utilizadas con mayor frecuencia de cierto tipo de *tweets*. En vez de ser una métrica esta sección permite explorar la composición de los *tweets*. Se espera que el conjunto de palabras frecuentemente utilizadas pueda dar señales del tipo de marihuana que se está consumiendo.

La Tabla 6.13 muestra ordenadamente las 20 palabras más frecuentes en *tweets* de consumo. Se

Palabra	Frecuencia (%)
weed	38,71
marihuana	28,26
pito	7,70
caño	7,20
sacate uno	4,81
porro	4,06
yerba	1,58
hierba	1,09
cuete	1,00
cannabis	0,74
pitito	0,56
cogollo	0,43
faso	0,35
ganja	0,30
marijuana	0,30
canabis	0,28
paragua	0,27
prensao	0,22
mariguana	0,20
cogollito	0,19

Tabla 6.13: Presencia de palabras en *tweets* de consumo

puede aprecia gran variedad de términos y gran utilización de palabras informales para denominar a la droga o su forma de consumo. Se hallan palabras de uso general y otras correspondientes a tipos de marihuana. Las palabras "paragua" y "prensao" hacen referencia precisa a una de las alternativas contempladas por la Encuesta Nacional de Drogas. Por otro lado, algunas palabras como "cogollo", "weed" y "hierba" pueden considerarse como consumo de marihuana verde.

A modo de comparación, en la Tabla 6.14 se mencionan las 5 palabras más frecuentes en *tweets* de políticas. Se observa una gran diferencia, debido a un gran uso de palabras formales para definir a la marihuana. Es importante notar que sólo dos palabras están presentes en el 98,32 % de los casos. Estas palabras son "marihuana" y "cannabis". Esta evidencia explica de alguna forma el mayor rendimiento en la clasificación de *tweets* de políticas.

Palabra	Frecuencia (%)
marihuana	81,84
cannabis	16,48
mariguana	0,49
canabis	0,47
marijuana	0,18

Tabla 6.14: Presencia de palabras en *tweets* de políticas

Capítulo 7

Conclusiones

7.1. Conclusiones Generales

Este estudio propone la utilización de la información generada en Twitter para replicar un comportamiento en la población general. El funcionamiento contempla la combinación de varios algoritmos y procedimientos para obtener los resultados deseados. La aplicación permite extraer información de los usuarios de Twitter y el contenido relacionado con marihuana que ellos mismos crearon. Así mismo, faculta la clasificación de los *tweets* con respecto a varias categorías y el cálculo de polaridad. Además de esto, implementa un modelo de predicción individual de consumo de marihuana.

Uno de los mayores valores de la aplicación es que brinda la posibilidad de extraer información útil desde *tweets*, que directamente son textos, el ejemplo clásico de información no estructurada. El rendimiento de los clasificadores sobre texto es medianamente bueno, bordeando el 65 % de *Precision* para la clase buscada y 84 % ponderada. Pero se pueden apreciar diferencias con respecto a cada clasificación. La clasificación de políticas en *tweets* es claramente mejor, indicando que dependiendo del tema, la división entre clases es más ambigua o requiere más información del contexto.

En este trabajo se reconoce el valor de las relaciones entre usuarios de Twitter, ya que sin ellas disminuiría en gran medida el poder predictivo del clasificador de consumo de marihuana. Además reproduce resultados obtenidos en otros estudios realizados con redes sociales fuera del contexto virtual. Implicando que el tipo de relación pasa desapercibido o que las relaciones en Twitter son reflejo de las relaciones de contacto directo. Se destaca que el consumo de marihuana es mayormente predicho por declaraciones de consumo por parte de amigos que las propias. Dándole respaldo a los estudios que señalan que el comportamiento de un individuo es fuertemente afectado por los pares.

Fue evidenciado un desfase de dos años entre los valores predichos por la aplicación y los recolectados por la Encuesta Nacional de Drogas. Señalando que el comportamiento se ve reflejado de forma retardada en las redes sociales, porque requiere que los individuos viertan esta información en sus cuentas. Aún así los datos son generados frecuentemente, ya que la polaridad es reportada

diariamente, y el predictor tiene capacidad de determinar consumo a nivel individual.

Todo esto no sería posible sin los permisos concedidos por Twitter para acceder a la información. Si bien los casos de bloqueo de información por parte de los usuarios no son menores, el porcentaje que no lo hace brinda una gran cantidad de información para realizar el análisis. Con el tiempo Twitter podría implementar políticas tan restrictivas como las de Facebook.

Finalmente se destacan los beneficios de la programación modular, principalmente simplifica la documentación. De esta manera, la aplicación es percibida como un conjunto elementos con interacciones claras en vez de un cúmulo de código. También la creación de un sistema desde cero produce beneficios. Si bien es más compleja, posee una enorme flexibilidad para resolver problemas específicos.

7.2. Trabajo Futuro

Como trabajo futuro se plantean dos líneas de desarrollo. Primero, mejorar el rendimiento del clasificador de consumo de personas. Esto se puede hacer mediante la incorporación de variables que expliquen de mejor manera la varianza del comportamiento. Por ejemplo, se puede utilizar una técnica más refinada de conexiones, reflejando la intensidad de la relación. Finalmente, se propone replicar el estudio a otras drogas, especialmente las lícitas como el alcohol y el tabaco. Es probable que tengan mayor presencia en las redes sociales y la metodología no requiere modificaciones.

Bibliografía

[1] C. C. Aggarwal and C. Zhai, *Mining text data*. Springer Science & Business Media, 2012.

[2] M. M. Ali, A. Amialchuk, and D. S. Dwyer, "The social contagion effect of marijuana use among adolescents," *PloS one*, vol. 6, no. 1, p. e16183, 2011.

[3] J. G. Bachman, L. D. Johnson, and P. M. O'Malley, "Explaining recent increases in students' marijuana use: impacts of perceived risks and disapproval, 1976 through 1996." *American journal of public health*, vol. 88, no. 6, pp. 887–892, 1998.

[4] J. G. Bachman, L. D. Johnston, P. M. O'Malley, and R. H. Humphrey, "Explaining the recent decline in marijuana use: Differentiating the effects of perceived risks, disapproval, and general lifestyle factors," *Journal of Health and Social Behavior*, pp. 92–112, 1988.

[5] S. L. Bailey, R. L. Flewelling, and J. V. Rachal, "Predicting continued use of marijuana among adolescents: The relative influence of drug-specific and social context factors," *Journal of Health and Social Behavior*, pp. 51–65, 1992.

[6] J. A. Balazs and J. D. Velásquez, "Opinion mining and information fusion: A survey," *Information Fusion*, vol. 27, pp. 95–110, 2016.

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

[8] M. O. Bonn-Miller, M. J. Zvolensky, and A. Bernstein, "Marijuana use motives: Concurrent relations to frequency of past 30-day use and anxiety sensitivity among young adult marijuana smokers," *Addictive Behaviors*, vol. 32, no. 1, pp. 49–62, 2007.

[9] T. Boudreau, J. Tulach, and G. Wielenga, *Rich client programming: plugging into the netbeans^{TM} platform*. Prentice Hall Press, 2007.

[10] J. D. Buckner, M. O. Bonn-Miller, M. J. Zvolensky, and N. B. Schmidt, "Marijuana use motives and social anxiety among marijuana-using young adults," *Addictive Behaviors*, vol. 32, no. 10, pp. 2238–2252, 2007.

[11] M. Cerdá, M. Wall, K. M. Keyes, S. Galea, and D. Hasin, "Medical marijuana laws in 50 states: investigating the relationship between state legalization of medical marijuana and marijuana use, abuse and dependence," *Drug and alcohol dependence*, vol. 120, no. 1, pp. 22–27, 2012.

[12] M. Chary, N. Genes, A. McKenzie, and A. F. Manini, "Leveraging social networks for toxicovigilance," *Journal of Medical Toxicology*, vol. 9, no. 2, pp. 184–191, 2013.

[13] K. Chen and D. B. Kandel, "Predictors of cessation of marijuana use: an event history analysis," *Drug and alcohol dependence*, vol. 50, no. 2, pp. 109–121, 1998.

[14] G. R. P. Chile, *Propuesta para un Chile mejor*, tercera ed., P. A. A. Palet, Ed. Chile: Felicidad, 2013.

[15] M. J. Cleveland, M. E. Feinberg, D. E. Bontempo, and M. T. Greenberg, "The role of risk and protective factors in substance use across adolescence," *Journal of Adolescent Health*, vol. 43, no. 2, pp. 157–164, 2008.

[16] N. Comeau, S. H. Stewart, and P. Loba, "The relations of trait anxiety, anxiety sensitivity, and sensation seeking to adolescents' motivations for alcohol, cigarette, and marijuana use," *Addictive behaviors*, vol. 26, no. 6, pp. 803–825, 2001.

[17] 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.

[18] O. Corazza, S. Assi, S. Malekianragheb, M. N. Beni, I. Bigdeli, Z. Aslanpour, and F. Schifano, "Monitoring novel psychoactive substances allegedly offered online for sale in persian and arabic languages," *International Journal of Drug Policy*, vol. 25, no. 4, pp. 724–726, 2014.

[19] E. Costa, A. Lorena, A. Carvalho, and A. Freitas, "A review of performance evaluation measures for hierarchical classifiers," in *Evaluation Methods for Machine Learning II: papers from the AAAI-2007 Workshop*, 2007, pp. 1–6.

[20] I. P. Cvijikj and F. Michahelles, "Monitoring trends on facebook," in *Dependable, Autonomic and Secure Computing (DASC), 2011 IEEE Ninth International Conference on*. IEEE, 2011, pp. 895–902.

[21] S. de Información Regional de México y Fundación Chile 21, *Políticas de drogas en México y Chile: Estimación de costos económicos y sociales y de escenarios alternativos*, 2013.

[22] A. C. de la Lengua and A. C. de la Historia, *Diccionario de uso del español de Chile (DUECh)*. MN Editorial, 2010.

[23] M. de Salud del Gobierno de Chile, *Objetivos Sanitarios 2011-2020*, 2011.

[24] P. Deluca, Z. Davey, O. Corazza, L. Di Furia, M. Farre, L. H. Flesland, M. Mannonen, A. Majava, T. Peltoniemi, M. Pasinetti *et al.*, "Identifying emerging trends in recreational drug use; outcomes from the psychonaut web mapping project," *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, vol. 39, no. 2, pp. 221–226, 2012.

[25] C. J. Dorius, S. J. Bahr, J. P. Hoffmann, and E. L. Harmon, "Parenting practices as moderators of the relationship between peers and adolescent marijuana use," *Journal of Marriage and Family*, vol. 66, no. 1, pp. 163–178, 2004.

[26] S. C. Duncan, T. E. Duncan, and H. Hops, "Progressions of alcohol, cigarette, and marijuana use in adolescence," *Journal of behavioral medicine*, vol. 21, no. 4, pp. 375–388, 1998.

[27] J. D. Dupré Casanova, "Generación de una metodología de detección de website keyobjects basado en métricas de centralidad de teoría de grafos," 2014.

[28] P. L. Ellickson, S. C. Martino, and R. L. Collins, "Marijuana use from adolescence to young adulthood: multiple developmental trajectories and their associated outcomes." *Health Psychology*, vol. 23, no. 3, p. 299, 2004.

[29] S. T. Ennett, K. E. Bauman, A. Hussong, R. Faris, V. A. Foshee, L. Cai, and R. H. DuRant, "The peer context of adolescent substance use: Findings from social network analysis," *Journal of research on adolescence*, vol. 16, no. 2, pp. 159–186, 2006.

[30] M. Fernández, "El costo socioeconómico del consumo de drogas ilícitas en chile," *Revista CEPAL*, vol. 107, pp. 93–114, 2012.

[31] N. Fieulaine and F. Martinez, "Time under control: Time perspective and desire for control in substance use," *Addictive Behaviors*, vol. 35, no. 8, pp. 799–802, 2010.

[32] S. A. Fleary, R. W. Heffer, E. L. J. McKyer, and D. A. Newman, "Using the bioecological model to predict risk perception of marijuana use and reported marijuana use in adolescence," *Addictive behaviors*, vol. 35, no. 8, pp. 795–798, 2010.

[33] Q. Fu, A. C. Heath, K. K. Bucholz, E. Nelson, J. Goldberg, M. J. Lyons, W. R. True, T. Jacob, M. T. Tsuang, and S. A. Eisen, "Shared genetic risk of major depression, alcohol dependence, and marijuana dependence: contribution of antisocial personality disorder in men," *Archives of General Psychiatry*, vol. 59, no. 12, pp. 1125–1132, 2002.

[34] S. Galea, A. Nandi, and D. Vlahov, "The social epidemiology of substance use," *Epidemiologic reviews*, vol. 26, no. 1, pp. 36–52, 2004.

[35] N. D. Galvão and H. d. F. Marin, "Data mining: a literature review," *Acta Paulista de Enfermagem*, vol. 22, no. 5, pp. 686–690, 2009.

[36] I. J. Ginsberg and J. R. Greenley, "Competing theories of marijuana use: A longitudinal study," *Journal of Health and Social Behavior*, pp. 22–34, 1978.

[37] A. Golub and B. D. Johnson, "Variation in youthful risks of progression from alcohol and tobacco to marijuana and to hard drugs across generations." *American Journal of Public Health*, vol. 91, no. 2, p. 225, 2001.

[38] C. A. Gonçalves, C. T. Gonçalves, R. Camacho, and E. C. Oliveira, "The impact of preprocessing on the classification of medline documents." in *PRIS*, 2010, pp. 53–61.

[39] K. Guo, L. Shi, W. Ye, and X. Li, "A survey of internet public opinion mining," in *Progress in Informatics and Computing (PIC), 2014 International Conference on*. IEEE, 2014, pp. 173–179.

[40] S. E. Hampson, J. A. Andrews, and M. Barckley, "Childhood predictors of adolescent marijuana use: early sensation-seeking, deviant peer affiliation, and social images," *Addictive behaviors*, vol. 33, no. 9, pp. 1140–1147, 2008.

[41] W. B. Hansen, J. W. Graham, J. L. Sobel, D. R. Shelton, B. R. Flay, and C. A. Johnson, "The consistency of peer and parent influences on tobacco, alcohol, and marijuana use among young adolescents," *Journal of behavioral medicine*, vol. 10, no. 6, pp. 559–579, 1987.

[42] R. Hogan, D. Mankin, J. Conway, and S. Fox, "Personality correlates of undergraduate marijuana use." *Journal of Consulting and Clinical Psychology*, vol. 35, no. 1p1, p. 58, 1970.

[43] B. A. Huberman, D. M. Romero, and F. Wu, "Social networks that matter: Twitter under the microscope," *Available at SSRN 1313405*, 2008.

[44] R. Jessor, "Predicting time of onset of marijuana use: a developmental study of high school youth." *Journal of Consulting and Clinical Psychology*, vol. 44, no. 1, p. 125, 1976.

[45] R. Jessor, J. A. Chase, and J. E. Donovan, "Psychosocial correlates of marijuana use and problem drinking in a national sample of adolescents," *American Journal of Public Health*, vol. 70, no. 6, pp. 604–613, 1980.

[46] R. Jessor and S. L. Jessor, "A social psychology of marijuana use: longitudinal studies of high school and college youth." *Journal of Personality and Social Psychology*, vol. 26, no. 1, p. 1, 1973.

[47] D. B. Kandel and J. A. Logan, "Patterns of drug use from adolescence to young adulthood: I. periods of risk for initiation, continued use, and discontinuation." *American journal of public health*, vol. 74, no. 7, pp. 660–666, 1984.

[48] H. B. Kaplan, S. S. Martin, R. J. Johnson, and C. A. Robbins, "Escalation of marijuana use: Application of a general theory of deviant behavior," *Journal of health and social behavior*, pp. 44–61, 1986.

[49] S. Keshav, "How to read a paper," *ACM SIGCOMM Computer Communication Review*, vol. 37, no. 3, pp. 83–84, 2007.

[50] K. M. Keyes, J. E. Schulenberg, P. M. O'Malley, L. D. Johnston, J. G. Bachman, G. Li, and D. Hasin, "The social norms of birth cohorts and adolescent marijuana use in the united states, 1976–2007," *Addiction*, vol. 106, no. 10, pp. 1790–1800, 2011.

[51] J. R. Kilmer, S. B. Hunt, C. M. Lee, and C. Neighbors, "Marijuana use, risk perception, and consequences: Is perceived risk congruent with reality?" *Addictive behaviors*, vol. 32, no. 12, pp. 3026–3033, 2007.

[52] K. Kobus and D. B. Henry, "Interplay of network position and peer substance use in early adolescent cigarette, alcohol, and marijuana use," *The Journal of Early Adolescence*, 2009.

[53] A. N. Kopstein, R. M. Crum, D. D. Celentano, and S. S. Martin, "Sensation seeking needs among 8th and 11th graders: characteristics associated with cigarette and marijuana use,"

Drug and alcohol dependence, vol. 62, no. 3, pp. 195–203, 2001.

[54] V. Korde and C. N. Mahender, "Text classification and classifiers: A survey," *International Journal of Artificial Intelligence & Applications*, vol. 3, no. 2, p. 85, 2012.

[55] R. Kosterman, J. D. Hawkins, J. Guo, R. F. Catalano, and R. D. Abbott, "The dynamics of alcohol and marijuana initiation: patterns and predictors of first use in adolescence." *American Journal of Public Health*, vol. 90, no. 3, p. 360, 2000.

[56] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques," 2007.

[57] J. W. LaBrie, J. F. Hummer, and A. Lac, "Comparing injunctive marijuana use norms of salient reference groups among college student marijuana users and nonusers," *Addictive behaviors*, vol. 36, no. 7, pp. 717–720, 2011.

[58] C. M. Lee, C. Neighbors, and B. A. Woods, "Marijuana motives: Young adults' reasons for using marijuana," *Addictive behaviors*, vol. 32, no. 7, pp. 1384–1394, 2007.

[59] G. Lee, R. L. Akers, and M. J. Borg, "Social learning and structural factors in adolescent substance use," *W. Criminology Rev.*, vol. 5, p. 17, 2004.

[60] B. Liu, "Sentiment analysis and opinion mining," *Synthesis lectures on human language technologies*, vol. 5, no. 1, pp. 1–167, 2012.

[61] C. Lopez-Quintero and Y. Neumark, "Effects of risk perception of marijuana use on marijuana use and intentions to use among adolescents in bogotá, colombia," *Drug and alcohol dependence*, vol. 109, no. 1, pp. 65–72, 2010.

[62] D. R. Miles, M. B. van den Bree, A. E. Gupman, D. B. Newlin, M. D. Glantz, and R. W. Pickens, "A twin study on sensation seeking, risk taking behavior and marijuana use," *Drug and alcohol dependence*, vol. 62, no. 1, pp. 57–68, 2001.

[63] D. S. Miller and T. Q. Miller, "A test of socioeconomic status as a predictor of initial marijuana use," *Addictive Behaviors*, vol. 22, no. 4, pp. 479–489, 1997.

[64] R. Myers, C.-P. Chou, S. Sussman, L. Baezconde-Garbanati, H. Pachon, and T. W. Valente, "Acculturation and substance use: Social influence as a mediator among hispanic alternative high school youth," *Journal of health and social behavior*, vol. 50, no. 2, pp. 164–179, 2009.

[65] D.-P. Nguyen, R. Gravel, R. Trieschnigg, and T. Meder, ""how old do you think i am?.ª study of language and age in twitter," 2013.

[66] M. Olavarría, "Estudio nacional sobre costos humanos, sociales y económicos de las drogas en chile, 2006," *Santiago, Chile, Olavarría y Asociados*, 2009.

[67] S. K. Pal, V. Talwar, and P. Mitra, "Web mining in soft computing framework: relevance, state of the art and future directions," *Neural Networks, IEEE Transactions on*, vol. 13, no. 5, pp. 1163–1177, 2002.

[68] J. Petraitis, B. R. Flay, and T. Q. Miller, "Reviewing theories of adolescent substance use: organizing pieces in the puzzle." *Psychological bulletin*, vol. 117, no. 1, p. 67, 1995.

[69] J. R. Ramirez, W. D. Crano, R. Quist, M. Burgoon, E. M. Alvaro, and J. Grandpre, "Acculturation, familism, parental monitoring, and knowledge as predictors of marijuana and inhalant use in adolescents." *Psychology of Addictive Behaviors*, vol. 18, no. 1, p. 3, 2004.

[70] L. Sáez, "El lenguaje secreto de las drogas en chileyerbagoma, jale, neo y afines," 1995.

[71] J. E. Schulenberg, A. C. Merline, L. D. Johnston, P. M. O'Malley, J. G. Bachman, and V. B. Laetz, "Trajectories of marijuana use during the transition to adulthood: The big picture based on national panel data," *Journal of Drug Issues*, vol. 35, no. 2, pp. 255–280, 2005.

[72] SENDA, *Décimo Estudio Nacional de Drogas en Población Escolar*, 2012.

[73] ——, *Décimo Estudio Nacional de Drogas en la Población General*, 2013.

[74] J. Simons and K. Carey, "An affective and cognitive model of marijuana and alcohol problems," *Addictive Behaviors*, vol. 31, no. 9, pp. 1578–1592, 2006.

[75] J. Simons and K. B. Carey, "Attitudes toward marijuana use and drug-free experience: Relationships with behavior," *Addictive Behaviors*, vol. 25, no. 3, pp. 323–331, 2000.

[76] J. Simons, C. J. Correia, and K. B. Carey, "A comparison of motives for marijuana and alcohol use among experienced users," *Addictive behaviors*, vol. 25, no. 1, pp. 153–160, 2000.

[77] J. Simons, C. J. Correia, K. B. Carey, and B. E. Borsari, "Validating a five-factor marijuana motives measure: Relations with use, problems, and alcohol motives." *Journal of Counseling Psychology*, vol. 45, no. 3, p. 265, 1998.

[78] J. Simons, R. Gaher, C. Correia, C. Hansen, and M. Christopher, "An affective-motivational model of marijuana and alcohol problems among college students." *Psychology of Addictive Behaviors*, vol. 19, no. 3, p. 326, 2005.

[79] J. S. Simons and K. B. Carey, "Personal strivings and marijuana use initiation, frequency, and problems," *Addictive Behaviors*, vol. 28, no. 7, pp. 1311–1322, 2003.

[80] J. S. Simons, M. S. Christopher, M. N. Oliver, and E. J. Stanage, "A content analysis of personal strivings: Associations with substance use," *Addictive behaviors*, vol. 31, no. 7, pp. 1224–1230, 2006.

[81] A. W. Stacy, "Memory activation and expectancy as prospective predictors of alcohol and marijuana use." *Journal of abnormal psychology*, vol. 106, no. 1, p. 61, 1997.

[82] A. L. Stone, L. G. Becker, A. M. Huber, and R. F. Catalano, "Review of risk and protective factors of substance use and problem use in emerging adulthood," *Addictive behaviors*, vol. 37, no. 7, pp. 747–775, 2012.

[83] J. S. Tucker, P. L. Ellickson, M. Orlando, S. C. Martino, and D. J. Klein, "Substance use

trajectories from early adolescence to emerging adulthood: A comparison of smoking, binge drinking, and marijuana use," *Journal of Drug Issues*, vol. 35, no. 2, pp. 307–332, 2005.

[84] Twitter. (2014, Nov.) Twitter developers. [Online]. Available: https://dev.twitter.com/

[85] M. B. van den Bree and W. B. Pickworth, "Risk factors predicting changes in marijuana involvement in teenagers," *Archives of general psychiatry*, vol. 62, no. 3, pp. 311–319, 2005.

[86] M. J. Van Ryzin, G. M. Fosco, and T. J. Dishion, "Family and peer predictors of substance use from early adolescence to early adulthood: An 11-year prospective analysis," *Addictive behaviors*, vol. 37, no. 12, pp. 1314–1324, 2012.

[87] D. Vlahov, S. Galea, H. Resnick, J. Ahern, J. A. Boscarino, M. Bucuvalas, J. Gold, and D. Kilpatrick, "Increased use of cigarettes, alcohol, and marijuana among manhattan, new york, residents after the september 11th terrorist attacks," *American journal of epidemiology*, vol. 155, no. 11, pp. 988–996, 2002.

[88] W3C. (2014, Nov.) World wide web consortium (w3c) web site. [Online]. Available: https://www.w3.org/WWW/

[89] F. Wagner and J. Anthony, "Into the world of illegal drug use: exposure opportunity and other mechanisms linking the use of alcohol, tobacco, marijuana, and cocaine," *American Journal of Epidemiology*, vol. 155, no. 10, pp. 918–925, 2002.

[90] F. A. Wagner and J. C. Anthony, "From first drug use to drug dependence: developmental periods of risk for dependence upon marijuana, cocaine, and alcohol," 2002.

[91] S. L. Wenzel, J. S. Tucker, D. Golinelli, H. D. Green, and A. Zhou, "Personal network correlates of alcohol, cigarette, and marijuana use among homeless youth," *Drug and alcohol dependence*, vol. 112, no. 1, pp. 140–149, 2010.

[92] H. R. White, B. J. McMorris, R. F. Catalano, C. B. Fleming, K. P. Haggerty, and R. D. Abbott, "Increases in alcohol and marijuana use during the transition out of high school into emerging adulthood: The effects of leaving home, going to college, and high school protective factors," *Journal of studies on alcohol*, vol. 67, no. 6, p. 810, 2006.

[93] J. Williams, R. Liccardo Pacula, F. J. Chaloupka, and H. Wechsler, "Alcohol and marijuana use among college students: economic complements or substitutes?" *Health economics*, vol. 13, no. 9, pp. 825–843, 2004.

[94] D. H. Wolpert, "The lack of a priori distinctions between learning algorithms," *Neural computation*, vol. 8, no. 7, pp. 1341–1390, 1996.

[95] L. L. Wright and T. P. Palfai, "Life goal appraisal and marijuana use among college students," *Addictive behaviors*, vol. 37, no. 7, pp. 797–802, 2012.

[96] K. Yamaguchi and D. B. Kandel, "On the resolution of role incompatibility: A life event history analysis of family roles and marijuana use," *American journal of Sociology*, pp. 1284–1325, 1985.

[97] J. S. Zeiger, B. C. Haberstick, R. P. Corley, M. A. Ehringer, T. J. Crowley, J. K. Hewitt, C. J. Hopfer, M. C. Stallings, S. E. Young, and S. H. Rhee, "Subjective effects to marijuana associated with marijuana use in community and clinical subjects," *Drug and alcohol dependence*, vol. 109, no. 1, pp. 161–166, 2010.

[98] M. J. Zvolensky, A. A. Vujanovic, A. Bernstein, M. O. Bonn-Miller, E. C. Marshall, and T. M. Leyro, "Marijuana use motives: A confirmatory test and evaluation among young adult marijuana users," *Addictive Behaviors*, vol. 32, no. 12, pp. 3122–3130, 2007.

Anexo A

Proceso de validación de ingreso

Ver página siguiente.

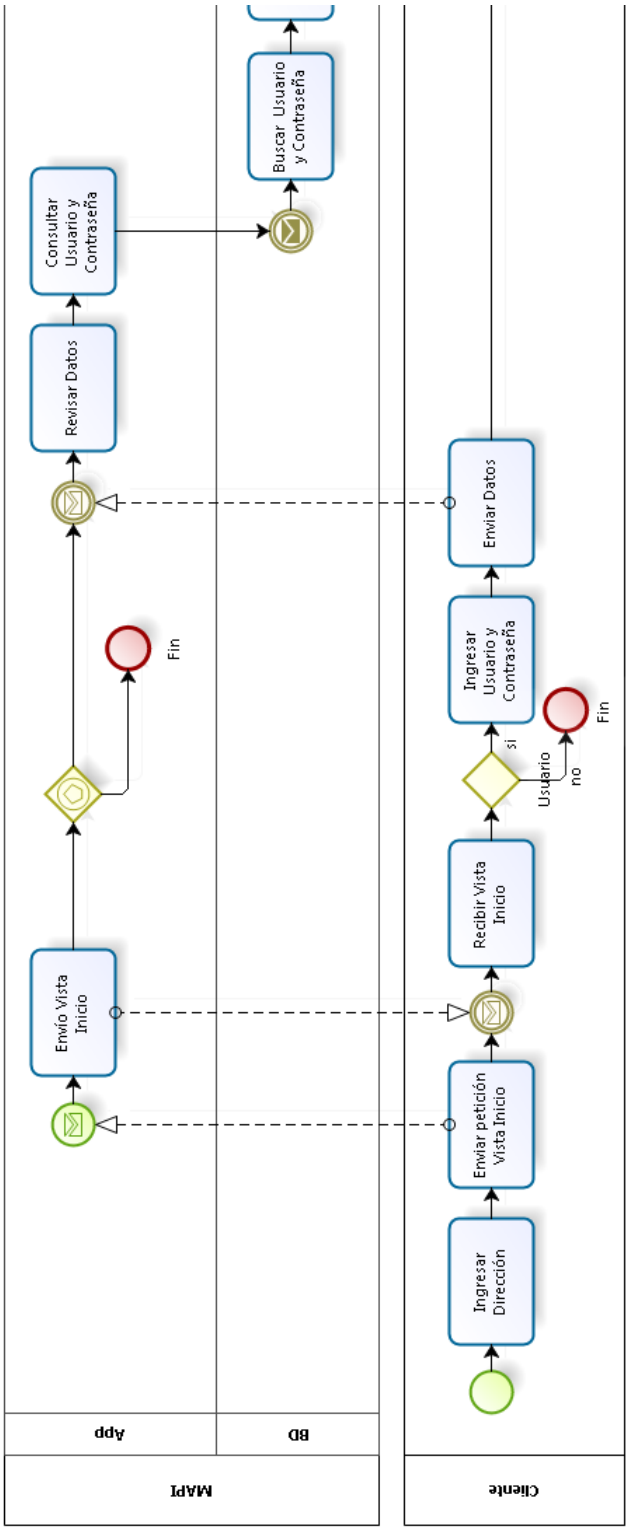

Figura A.1: Proceso BPMN de ingreso (1)
Fuente: Elaboración Propia

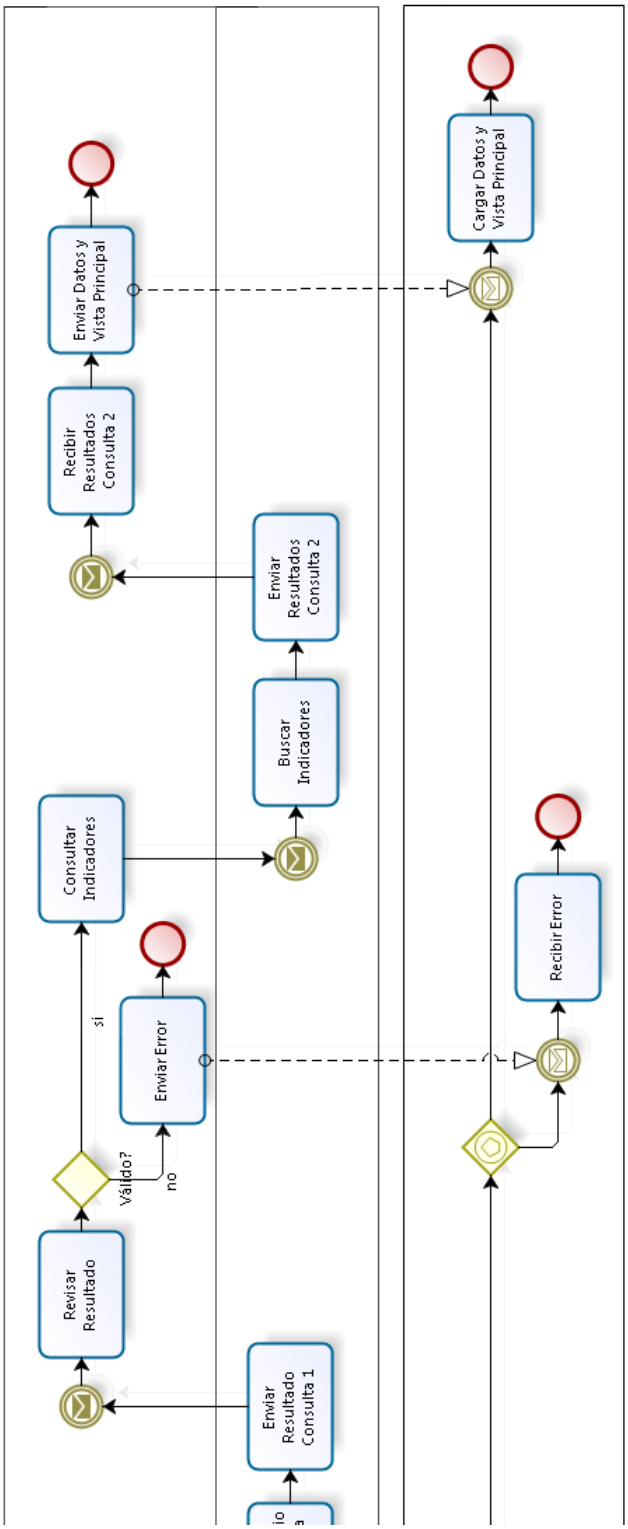

Figura A.2: Proceso BPMN de ingreso (2)
Fuente: Elaboración Propia

Anexo B

Vistas de prototipo funcional

Figura B.1: Vista de login
Fuente: Elaboración Propia

Figura B.2: Vista de polaridad anual
Fuente: Elaboración Propia

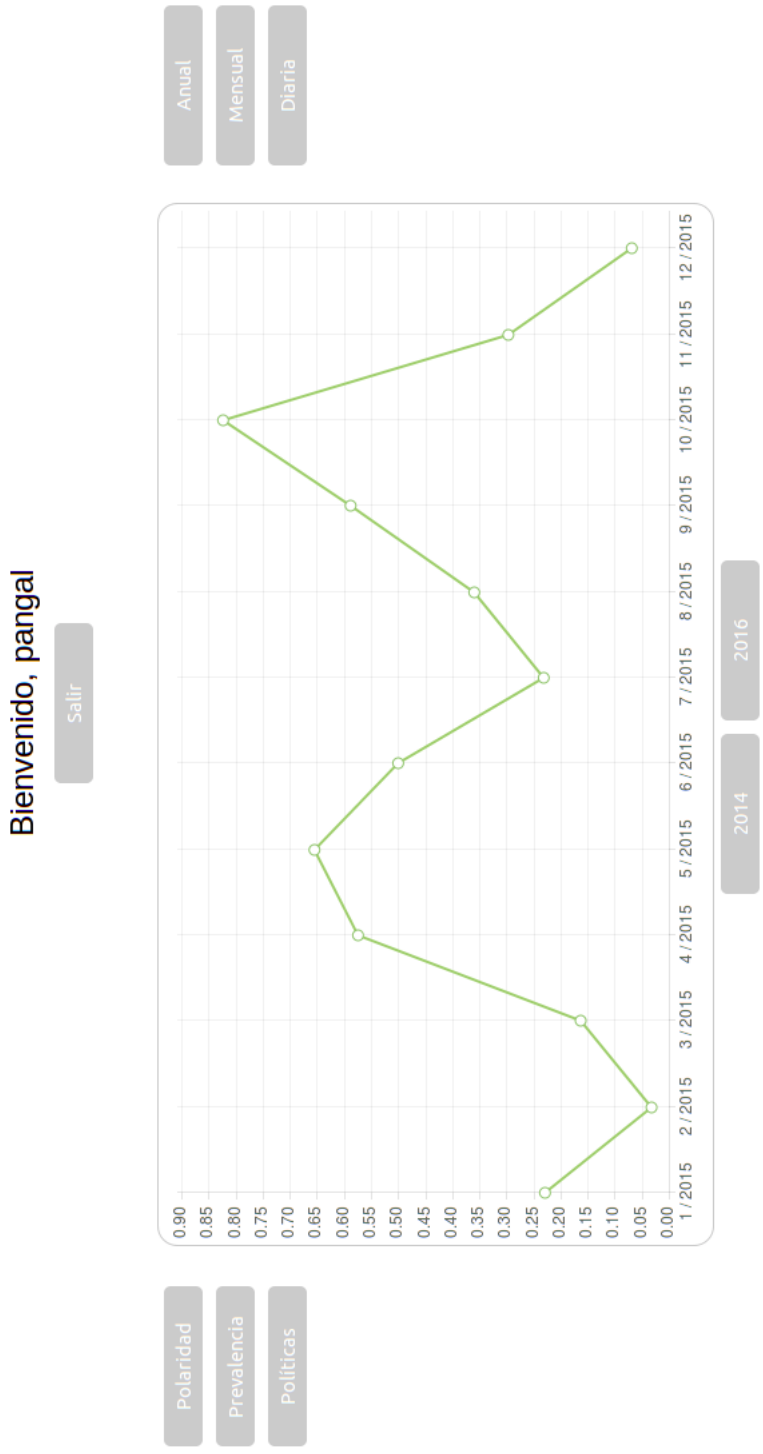

Figura B.3: Vista de polaridad mensual
Fuente: Elaboración Propia

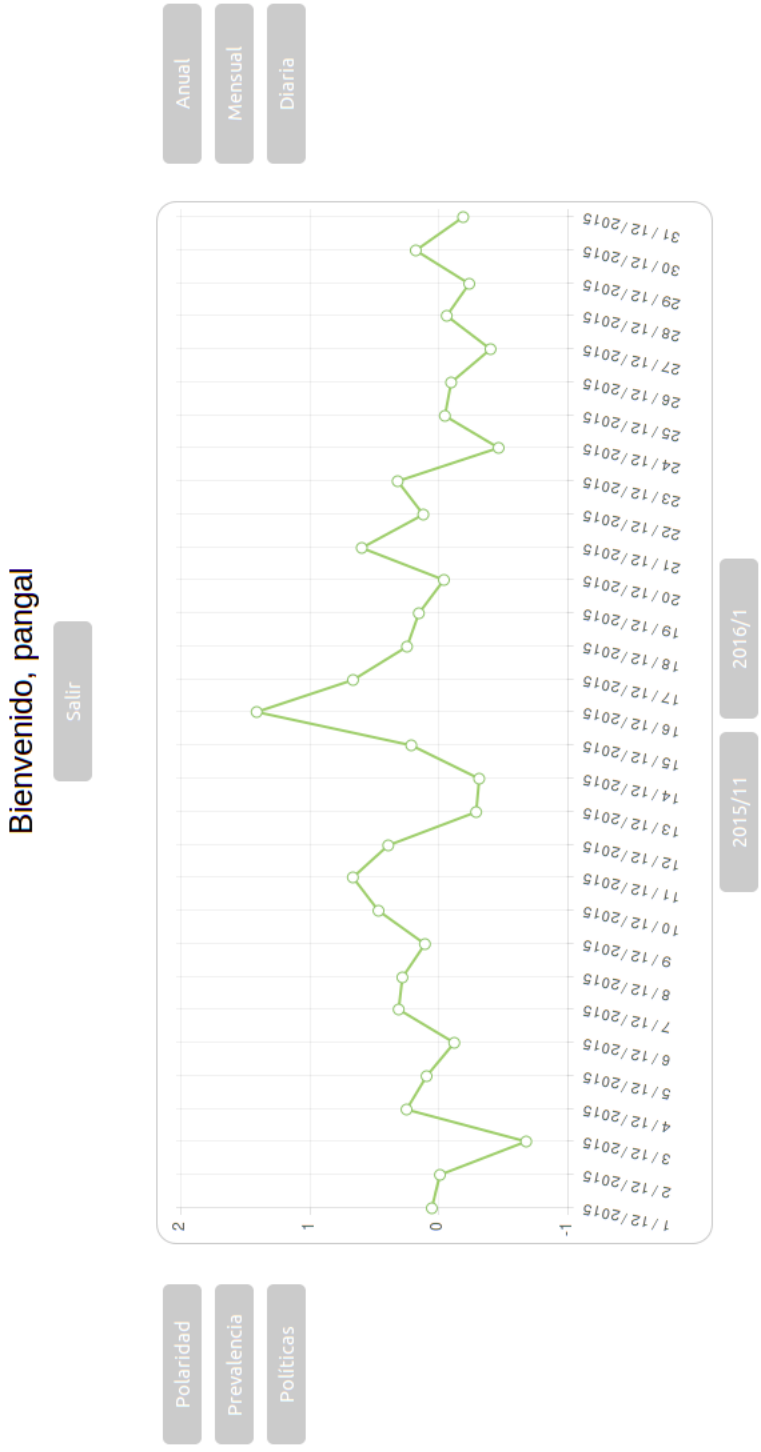

Figura B.4: Vista de polaridad diaria
Fuente: Elaboración Propia

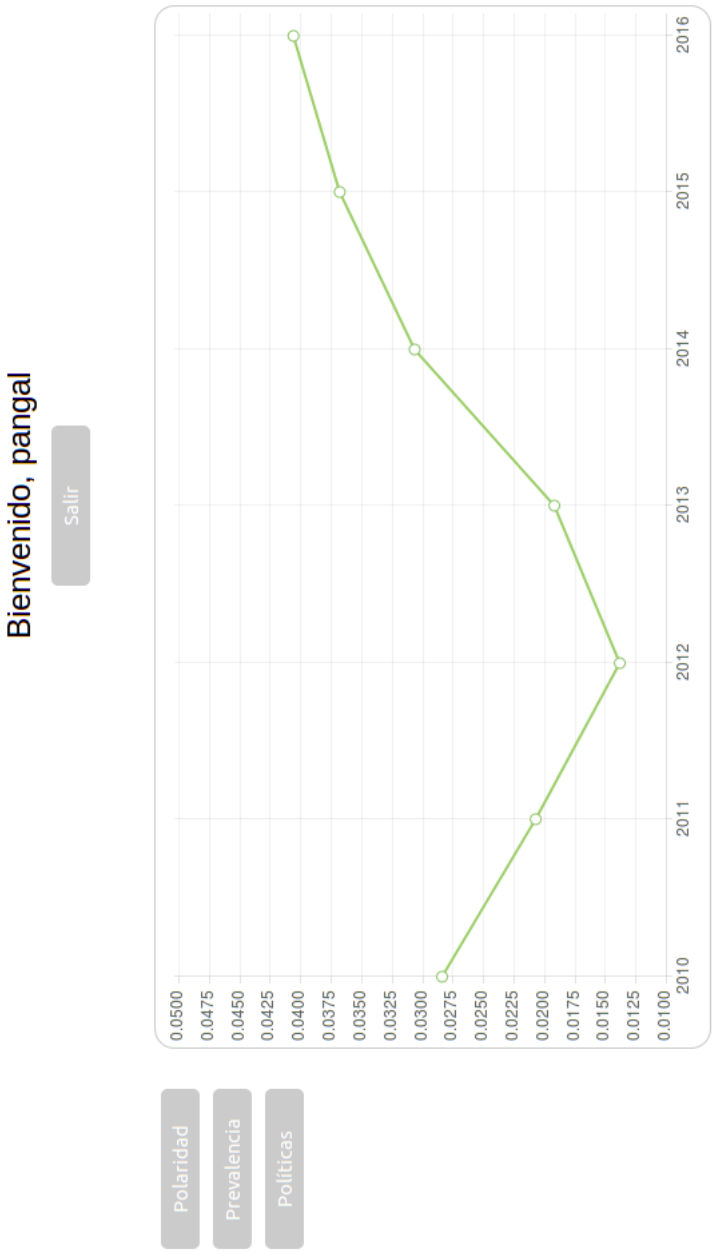

Figura B.5: Vista de prevalencia
Fuente: Elaboración Propia

Figura B.6: Vista de polaridad de políticas
Fuente: Elaboración Propia