Desentrañando Twitter para replicar los indicadores de prevalencia y percepción de riesgo de la marihuana utilizados en el Estudio Nacional de Drogas de Chile

Víctor D. Cortés [*]
Felipe E. Vildoso [*]
Juan D. Velásquez [*]
Carlos F. Ibáñez [**]

Resumen

El objetivo de este trabajo es diseñar y desarrollar un sistema que recoja información de Twitter con el fin de monitorizar el consumo de marihuana y su percepción de riesgo. La clasificación de texto y sentimientos emitida por los usuarios, y las conexiones entre usuarios (entorno social) son utilizados para construir indicadores de consumo y percepciones con respecto a la marihuana. Se obtuvo un conjunto de 1,361,285 usuarios chilenos. Los clasificadores de texto y consumo de marihuana individual tuvieron medidas de precisión superiores a 0,7. Se concluye que es posible construir un sistema que utilice a Twitter como fuente de datos para reproducir tendencias con respecto a la marihuana a nivel individual y agregado. Esta información permitiría complementar los resultados de los estudios nacionales de drogas y contribuir a informar las políticas de drogas en el país.

Palabras Clave: *Marihuana*, *Prevalencia*, *Percepción de Riesgo*, *Aprendizaje de Máquinas*, *Redes Sociales*.

[*]Departamento de Ingeniería Industrial, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile, Santiago, Chile.
[**]Departamento de Psiquiatría y Salud Mental Norte, Facultad de Medicina, Universidad de Chile, Santiago, Chile.

1. Introducción

Desde inicios del siglo XXI se ha evidenciado un incremento abrupto en el uso y penetración de Internet. La introducción de la Web 2.0 hizo posible la generación de contenido por parte de usuarios que antes estaban restringidos a la lectura, facilitando el flujo de ideas y conocimiento por medio de contenido informal. Este evento fue acompañado de un crecimiento sostenido por varios años del número de usuarios. La irrupción de las redes sociales no vino más que a acentuar esta evolución, ya que hizo aún más evidente el cambio de paradigma.

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 a sumergirse en estos ambientes. Por esta razón, estos escenarios son interesantes desde el punto de vista de recolección de información que puede representar a las personas y sus comportamientos [**?**].

Por otro lado, a lo largo de la historia, el consumo de drogas ha sido asociado a 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 [7] y organismos privados.

En Chile, la marihuana es un caso especial de estudio. Esta droga tiene atención particular, debido al amplio debate y relevancia que se le ha dado. En los últimos años, la percepción de riesgo de la droga ha disminuido, reflejando una norma social a favor del consumo y pudiendo ser origen del alza significativa del consumo nacional. En efecto, para el año 2012, la tasa anual de prevalencia fue de 7,1 %, 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.

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 y

percepción de riesgo, entre otras variables.

El costo del Estudio Nacional de Drogas en sus versiones de población general y escolar hacen que el periodo entre estudios (cada dos años) sea mayor al 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.

En este contexto, la aplicación de técnicas de minado de datos pueden traducirse en ventajas significativas al momento de procesar información estructurada y en forma de textos, para brindar datos complementarios a los mecanismos usados actualmente. La dificultad de esta iniciativa reside en estructurar gran cantidad de texto e información relativa a los usuarios y sus interacciones. Dichos datos tienen que ser transformados en información que apoye a la toma decisiones [**?**].

En síntesis, este estudio busca replicar los indicadores de prevalencia y percepción de riesgo de la marihuana del Estudio Nacional de Drogas. La similitud de las redes sociales, específicamente Twitter, con la forma natural en que se relacionan las personas y la variedad de información vertida en ellas, permite plantear 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*". Para validarla, fue recolectado un conjunto de usuarios chilenos de Twitter, su información personal, sus relaciones de seguimiento y sus tweets. En base a estos datos fueron calculadas métricas de Análisis de Redes Sociales, fueron aplicadas técnicas de minería de datos para clasificar textos, se utilizaron algoritmos de Opinion Mining para determinar sentimientos en los textos, y posteriormente se obtuvieron patrones con relación al consumo de marihuana.

El paper está estructurado como se describe a continuación: la Sección 2 hace revisión de algunos estudios de métricas utilizadas para explicar el consumo de marihuana y otros explorando las redes sociales online en relación a ciertas drogas. La Sección 3 muestra detalladamente nuestra propuesta, estableciendo las métricas y los modelos que fueron utilizados para validar la hipótesis. En la Sección 4 se describe la metodología de recolección de datos. Luego, la Sección 5 contiene detalles correspondiente al tratamiento de datos y a los resultados obtenidos. Finalmente, la Sección 6 presenta las conclusiones finales y posibles líneas de trabajo futuro.

2. Trabajo relacionado

En esta sección serán mencionadas varias lineas de investigación con respecto al consumo y opinión sobre la marihuana. En primer lugar, serán descritos estudios que evaluaron factores protectores y de riesgo con respecto al consumo. Estos estudios generalmente exploran la relación directa o indirecta de un conjunto de variables explicativas sobre el comportamiento observado. En segundo lugar, serán detallados estudios relacionados con Análisis de Redes Sociales para explicar el consumo de marihuana, ya que son los más acordes con la propuesta de investigación. En tercer lugar, serán nombrados algunos estudios relacionados con la exploración de drogas en entornos web. En último lugar, se hará una revisión de estudios aplicando Opinion Mining.

Varios han sido los intentos por establecer una teoría que explique el consumo de substancias. Todas ellas, al igual que las variables que utilizan, tienen elementos en común, algunos elementos diferenciadores, y diferentes niveles de respaldo empírico. [15] es un intento por entender las similitudes, diferencias, intersecciones y vacíos de las distintas teorías más prominentes. En él son reconocidos varios niveles de influencia en el consumo de marihuana. Estos niveles son tres: intrapersonal, cultural e interpersonal. Este último está enfocado en el efecto que tiene el contorno social sobre el consumo individual de substancias.

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 [14], las mayoría de las variables propuestas del modelo se sostienen en un modelo estable de predicción de consumo de marihuana. Estas variables presentaron efectos directos y mediaron otro tipo de variables estructurales como el género, clase o edad.

En [18] y [5] 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 se comparaban varias teorías y se indagaba el efecto conjunto de variables grupales y psicológicas concluyeron que la orientación hacia un grupo de referencia usuario de marihuana es el predictor más sustancial de uso de marihuana ([9]).

En [20] se señaló que jóvenes con más usuarios de sustancias en sus redes sociales reportaron mayor consumo. Más precisamente, en [1], 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 %. [19] agregó 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 [17], 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.

El Análisis de Redes Sociales consiste en construir un conjunto de métricas a partir de las conexiones entre las personas. Es posible construir redes complejas en base a cinco nominaciones de amigos por cada uno de los sujetos encuestados. A partir de esto, la posición dentro de la estructura de la red social también influye en la conducta. Efectivamente, [8] 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 [12] se halló que el consumo 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.

Las normas sociales son ligadas con la percepción del individuo de la aprobación de los pares sobre algún comportamiento. Aplicado a la marihuana, en [11] fueron evidenciadas variaciones de desaprobación de la marihuana en distintos cortes generacionales y que estas diferencias afectaron directamente 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, [13] 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.

Las comunidades online también fueron exploradas en relación al consumo de drogas. En [16] fue examinada la conexión entre las características de la red online y el consumo de sustancias de adultos jóvenes. El uso de drogas fue asociado con un elevado número de conexiones, y una elevada proporción de la red que discute y acepta el consumo de drogas. También se halló que la densidad de la red y el número total de conexiones fueron asociados a mayor consumo personal en hombres.

En [6] se evaluó la asociación entre la presencia del contenido de consumo de alcohol y otras drogas, las normas percibidas, y el consumo de marihuana en adultos jóvenes. El remordimiento anticipado fue negativamente asociado con el consumo de marihuana. Al igual que en estudios mencionados antes, el consumo por parte de pares fue positivamente asociado con el consumo individual. Esto quiere decir que los resultados son similares tanto para efectos

sociales online como efectos sociales normales.

La idea de explorar las redes sociales con el fin de complementar medios tradicionales de recolección de información con respecto a drogas no es nueva. En [4] se introducen herramientas para usar datos desde redes sociales. Se sugiere un enfoque estructurado para capturar tendencias emergentes en el abuso de drogas aplicando métodos de inteligencia artificial, computación lingüística, teoría de grafos y modelamiento basado en agentes. Es más, se sugiere a Twitter como una red social disponible públicamente para obtener datos.

Otra forma usada para obtener información de las redes sociales es la aplicación de técnicas de Opinion Mining. Este es un sub-campo de Text Mining que permite extraer las opiniones desde documentos. Específicamente, puede ser aplicado en el contenido emitido por los usuarios de la Web. [3] hace una revisión bibliográfica reciente con respecto al sub-campo. En él, se plantea el problema de extracción de opiniones y se nombran los principales enfoques para resolverlo. [2] detalla la utilización del enfoque no supervisado basado en lexicones. Este enfoque explota reglas y heurísticas obtenidas del lenguaje, basándose en la polaridad individual de las palabras y la aplicación de reglas que pueden cambiar o intensificar la polaridad del conjunto de palabras. Particularmente, este trabajo tuvo como resultado una API de Opinion Mining.

3. Propuesta de Investigación

El objetivo de este estudio es replicar los indicadores de prevalencia y percepción de riesgo de la marihuana del Estudio Nacional de Drogas, utilizando datos recolectados únicamente desde Twitter. La idea inicial es construir una red social de usuarios chilenos de Twitter con información suficiente para reconocer patrones en el comportamiento de las personas.

3.1. Hipótesis de Investigación

La forma en que los usuarios de las redes sociales generan información y el modo en que éstos se relacionan, incentivan a entidades a sumergirse dentro de esos escenarios para obtener conocimiento. Especialmente, la semejanza entre la estructura de las redes reales y online, y el tamaño de las comunidades promueven el pensamiento de que cualquier descubrimiento obtenido allí es generalizable para la población. A todo esto, es sumado el estado actual de algoritmos que facilitan el cálculo de sentimientos y clasificación de un texto, un tipo de dato sin estructura.

Basado en lo expuesto antes, se declara la siguiente hipótesis de investigación: *"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"

En este sentido, se quiere clasificar a los textos (*tweets*) y calcular su polaridad de sentimientos para luego transformarlos en variables asociadas a cada usuario. En base a éstas y a otras métricas derivadas de Análisis de Redes Sociales se pretende clasificar el consumo de marihuana a nivel individual. Además se quiere evaluar la opinión con respecto a la droga. Estos resultados pueden utilizarse en la elaboración de índices agregados que ayuden a explicar el consumo de marihuana a nivel nacional.

3.2. Modelos

En primer lugar, se requiere calcular la polaridad de sentimientos para los *tweets*, y clasificar a los mismos 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).

Cada punto anterior se llevará a cabo mediante algoritmos de aprendizaje supervisado, con excepción del cálculo de polaridad de sentimientos, ya que éste se implementará mediante algoritmos de Opinion Mining, utilizando el enfoque no supervisado basado en lexicón. En los textos, las variables generalmente nacen desde las mismas palabras representadas de forma matricial. Esta representación se utilizará para las tres clasificaciones de *tweets* y la edad en usuarios. Luego de aplicar esta transformación, el problema se convierte en uno típico de minería de datos. Por lo tanto, junto con el consumo en usuarios, serán evaluados los algoritmos que suelen obtener mejor rendimiento. Por ejemplo, Redes Neuronales, Support Vector Machines, Naïve Bayes, etc.

En la evaluación de rendimiento para cada modelo de clasificación, serán utilizadas algunas métricas derivadas de la matriz de confusión: *Precision*, *Recall* y *F-Measure*. Priorizando la primera por sobre las demás. Además será utilizada la técnica de validación cruzada en el entrenamiento de algoritmos.

4. Recolección de Datos

En esta sección se describirá el tipo de información que fue utilizada como datos de entrada para el estudio. 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. Más tarde, se precisará la forma para obtener los datos desde su fuente. Finalmente, se definirán por separado los mecanismos de etiquetado de textos y de usuarios para sus respectivos análisis y algoritmos de clasificación.

4.1. Datos disponibles

La información disponible en *Twitter* está moldeada por las funcionalidades que ofrece su servicio de microblogging. Para efectos de este estudio, 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.

4.2. Estructura Necesaria

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 1 y 2 muestran ejemplos de los etiquetados necesarios para *tweets* y para usuarios, respectivamente.

4.3. Recolección de Datos

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. El objetivo fue 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.

El algoritmo de obtención de usuarios operó 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

	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 1: 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 2: Etiquetado de usuarios

todos los elementos adyacentes a él. Este procedimiento fue 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 este estudio, el criterio consiste en que los usuarios sean chilenos, cuya información está contenida en los datos del usuario.

En la Figura 1 se muestra el proceso iterativo de extracción de usuarios y en la Figura 2 se muestra un ejemplo de dos iteraciones. El algoritmo comienza con una semilla en la Figura 2.a, en las figuras 2.b y 2.d se incorporan los nodos adyacentes que cumplen el criterio (nodos verdes) y en las figuras 2.c y 2.e se muestra el estado final de cada iteración.

Es importante mencionar que existe un porcentaje de usuarios de *Twitter* que bloquean el acceso a sus *tweets*. Para efectos del diseño, no fueron considerados los usuarios con esta condición.

Una vez establecida la base de usuarios se procedió a extraer el conjunto de *tweets* publicado por cada uno. El modo de clasificar los *tweets* relacionados con marihuana consistió en identificar palabras clave. El listado de palabras clave tuvo como origen a tres fuentes diferentes: conocimiento experto, bibliografía y una encuesta de uso de palabras actuales.

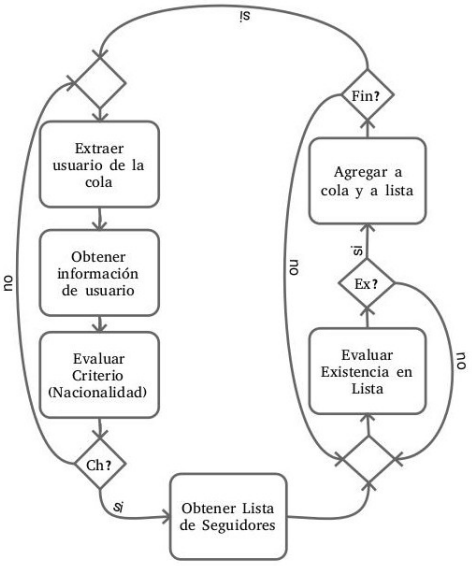

Figura 1: Crawler de Usuarios

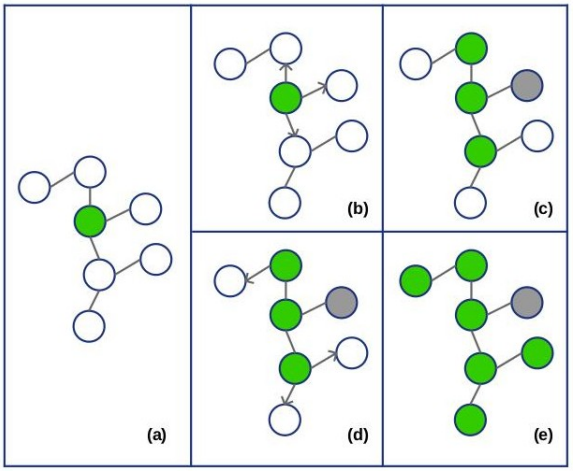

Figura 2: Ejemplo de Iteración

El conjunto total de palabras fue filtrado para confirmar su uso en *Twitter* y desambiguar el contexto de uso. Para hacer esto se extrajo un grupo *tweets* que contiene las palabras. Fueron aplicados algunos algoritmos de preprocesamiento de texto para emplear *Topic Modeling* con el fin de identificar contextos diferentes de uso de las palabras (clusters) y verificar su empleo relacionado con marihuana. Luego, las palabras ambiguas fueron filtradas, considerando sólo aquellos *tweets* que contuvieran 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.

Una vez que un *tweet* ha pasado el filtro, se calcula su polaridad con una API desarrollada en [2] que aplica Sentiment Analysis con el enfoque no supervisado basado en un lexicón etiquetado por emoción y un puntaje para cada término. El lexicón permite obtener un puntaje que va desde -50 a +50 donde el positivo indica que un texto habla positivamente, mientras que un puntaje negativo indica que la emoción es negativa. El lexicón utilizado fue desarrollado en [10]. Esta API aplica tres reglas gramaticales del lenguaje natural del español. La primera regla consiste en los intensificadores, es decir, palabras que amplifican el significado de una palabra que se encuentre en el lexicón. La segunda regla, es la de la negación, la cual invierte el valor que se obtiene en el lexicón. La tercera, y última regla, corresponde a las clausulas adversativas, esto es el uso de "pero" o palabras similares, en donde la primera parte antes de la clausula recibe una ponderación menor a la que viene después.

4.4. Etiquetado de *Tweets*

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.

Dada la experiencia del grupo investigador, se buscó un número de *tweets* para la muestra que cumpliera con tener un error del 3 % al 98 % de confianza.

Teniendo esos requerimientos se hace necesario contar con una colección de *tweets* de al menos 1.500. Este conjunto de *tweets* fue clasificado por un grupo de 12 personas. Es importante mencionar que no es sencillo contar con personas que pueda etiquetar manualmente todos los textos, por lo que se utilizo la siguiente metodología para que no tuvieran que etiquetar todos los datos. Primero, se separaron 50 tweets de los 1.500, los cuales fueron etiquetados por las 12 personas y que sirve para poder ver la concordancia entre los participantes. Luego, de los 1.450 restantes, se dividieron en 12 grupos para que cada evaluador se encargara solamente de 1 de ellos. Por último, la división y la distribución de textos fue hecha al azar.

4.5. Etiquetado de Usuarios

El etiquetado se desarrolló mediante una encuesta directa a los usuarios. Una vez construida una base de datos de usuarios chilenos, se escogieron casos al azar para enviarles la encuesta. La cual fue publicada en un *tweet* para cada usuario, mencionándole directamente. La encuesta contuvo preguntas para determinar el consumo de marihuana, la edad y el sexo de cada usuario.

5. Resultados y Discusión

En esta sección será presentado el cúmulo de resultados derivados del estudio. 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 fueron utilizadas en la recolección de *tweets*. A continuación de esto, se hará referencia a información originada en la recolección de datos. Luego, serán mencionados algunos datos que fueron obtenidos en el etiquetado de *tweets* y usuarios. La evaluación de algoritmos también tendrá destinada un segmento. Finalmente, se pondrá enfoque en las métricas elaboradas a partir de los resultados anteriores.

5.1. Palabras Clave

El proceso de selección de palabras 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 3.

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 3: Palabras Clave

5.2. Recolección de Datos de *Twitter*

La velocidad de recolección de *tweets* y usuarios fue una medida crucial durante el estudio. 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 influyó en la cantidad de usuarios que fueron evaluados y almacenados cada día.

La Figura 3 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 3.

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 4 y 5, 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 4 revela que para años anteriores al 2009 existían poco usuarios, por lo que desacredita resultados que puedan ser originados para esos años. La Figura 5 muestra la composición de *tweets* en la base de datos, cuya forma está determinada por la restricción de *Twitter* de los últimos 3.200 *tweets* por

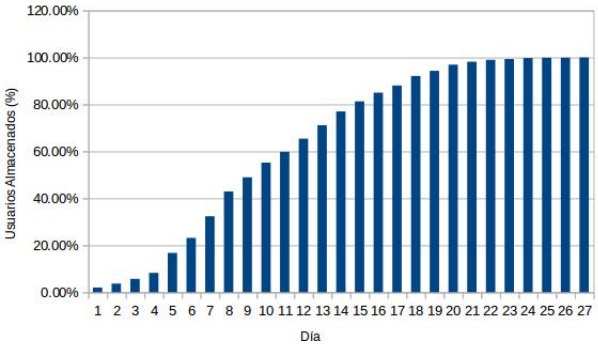

Figura 3: Gráfico de usuarios acumulados

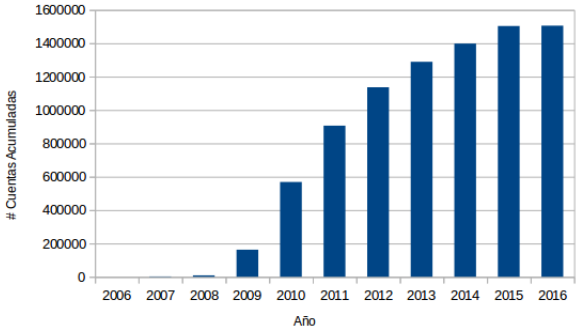

Figura 4: Gráfico de cuentas acumuladas

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 de 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 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.

5.3. Etiquetado de *Tweets*

A lo largo de los capítulos se han mencionado las categorías en que era necesario etiquetar los *tweets*. Tres de éstas están dedicadas para el entrenamiento de algoritmos y otra para determinar la precisión de las palabras clave.

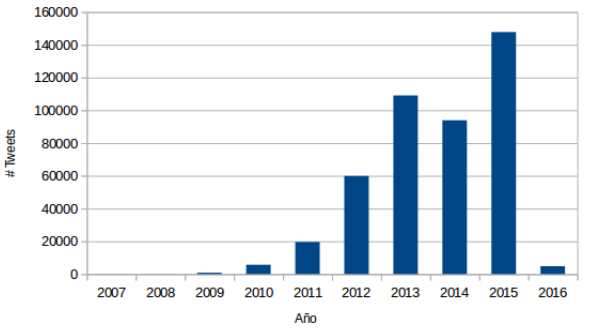

Figura 5: Número de *tweets* por año

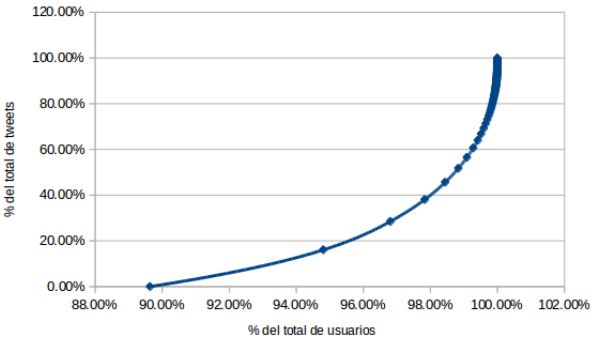

Figura 6: Curva de Lorenz de *Tweets*

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 4: Medidas de Acuerdo

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

Tabla 5: Heterogeneidad en las etiquetas

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 4 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 4, todas las categorías muestran un coeficiente cercano al 60 %, a excepción de 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 al promedio en cada uno de los 50. Los porcentajes para los casos positivamente clasificados se muestran en la Tabla 5. 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 de cualquier clasificador confiable. Por esta razón y su coeficiente *Kappa*, esta categoría es apartada de cualquier manipulación en etapas siguientes.

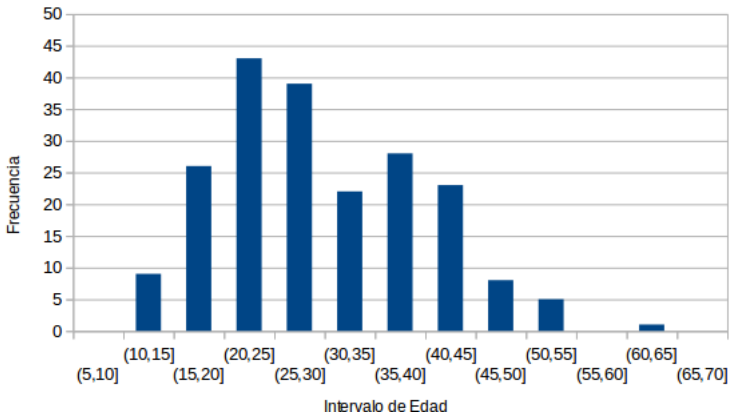

Figura 7: Distribución de edad de la muestra

5.4. Etiquetado de 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 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.

5.5. Evaluación de Algoritmos

En esta sección fueron elegidos los algoritmos que contaron con las mejores medidas de rendimiento para las tareas de clasificación y regresión. En cada tarea fueron evaluados varios algoritmos, fue seleccionado aquel que contó con mayor poder predictivo.

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

- 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 su procedimiento de evaluación no será detallado. El rendimiento de cada algoritmo incorpora intrínsecamente estas modificaciones, siendo sólo nombradas junto al algoritmo.
- Varios algoritmos de entrenamiento fueron descartados por sus 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 priorizó aquellos que tuvieron mayor precisión para la clase de interés. Esto refleja la necesidad de recuperar casos en que efectivamente se evidencie el comportamiento.

5.5.1. 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 fue evaluada 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, 7 y 8 muestran las medidas de rendimiento para cada uno de los algoritmos. En ellos se aprecian 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 priorizó la *Precision* de la clase de consumo, por lo tanto fue 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: 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 7: 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 8: Rendimiento de SVM para el consumo en *tweets*

5.5.2. 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 9, 10 y 11 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 residió 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 *Pre-*

cision. 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.

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 9: 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 10: 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 11: Rendimiento de C4.5 para políticas en *tweets*

5.5.3. 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 ocupan. En esta ocasión son utilizadas medidas de relación lineal y diferencias agregadas entre los datos reales y los predichos. Específicamente se utilizan la correlación de *Pearson* y otros errores.

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

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 12: Rendimiento de algoritmos de edad

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 13: Rendimiento de SVM para consumo en usuarios

de frecuencias log-normalizados. Aunque el mejor modelo tiene una pequeña variación. Las medidas de rendimiento se muestran en la Tabla 12, 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 promedio cerca de 6 años.

5.5.4. 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 a 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 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 13 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 14 muestra los pesos normalizados para cada variable. Los datos

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 14: Influencia de variables en el consumo de marihuana

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.

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. La popularidad de una persona aumenta 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.

5.6. Indicadores

En esta sección será mostrado el producto final de esta memoria: las réplicas inicialmente prometidas y resultados. 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.

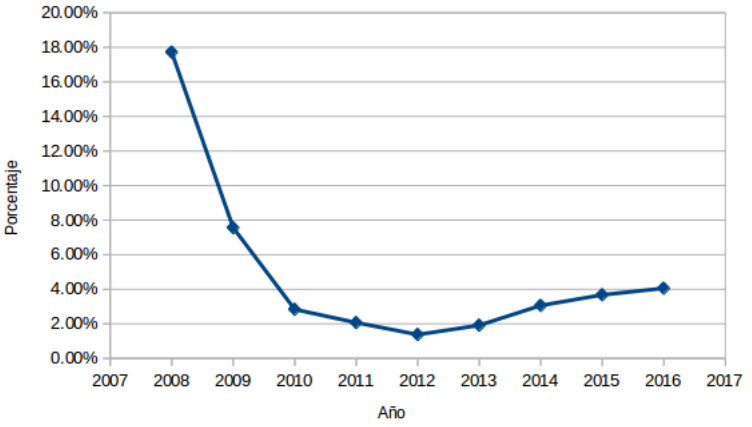

Figura 8: Prevalencia Anual

5.6.1. 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 8 muestra el cálculo de prevalencia para cada año, la cual revela un alto porcentaje para los 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 de 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 8 y los datos recogidos por la Encuesta Nacional de Drogas. La prevalencia histórica arrojada por esta encuesta es expuesta en la Figura 9. 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 10 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.

La gran similitud entre curvas es innegable, presentando un coeficiente de correlación de *Pearson* de 0,933. Aunque hay que destacar que la curva pre-

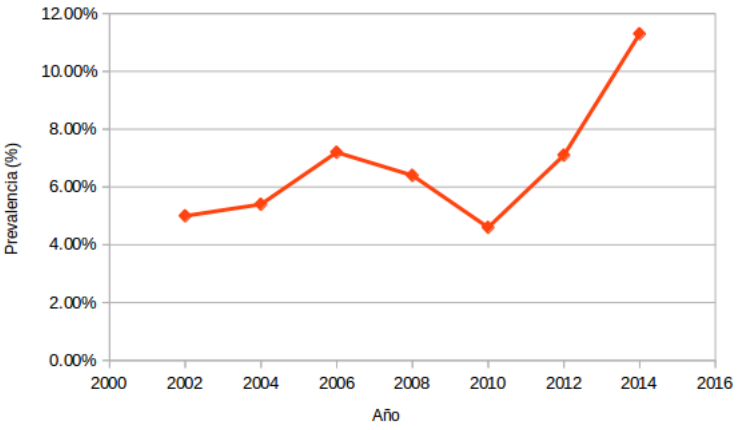

Figura 9: Prevalencia Nacional

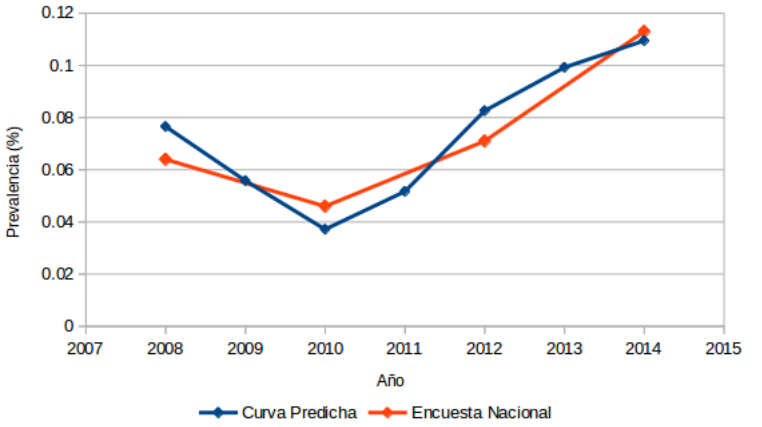

Figura 10: Comparación de Prevalencia

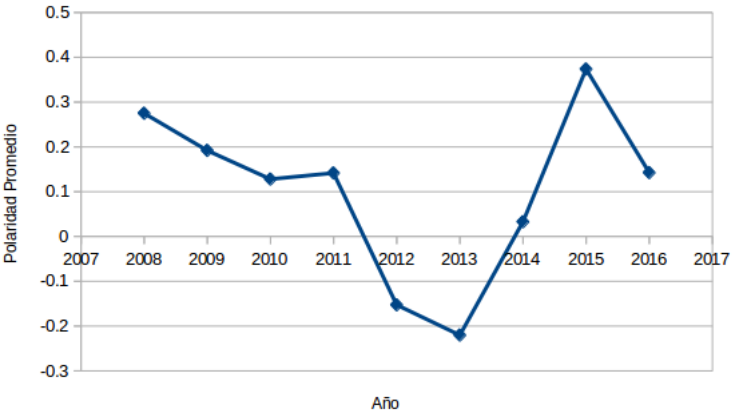

Figura 11: Polaridad de *Tweets*

dicha 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 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.

5.6.2. 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 11 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.

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 12 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

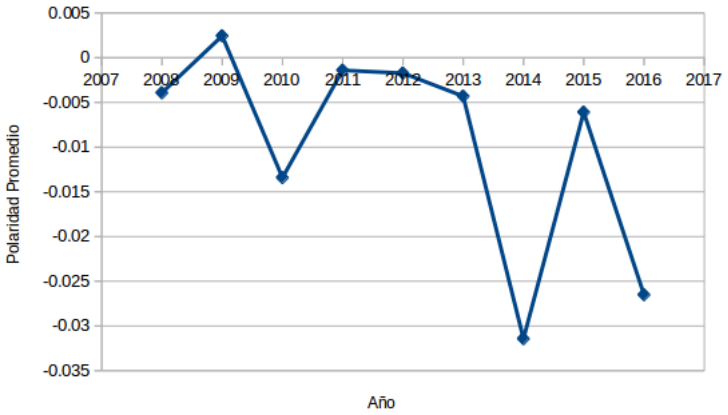

Figura 12: Polaridad de Usuarios

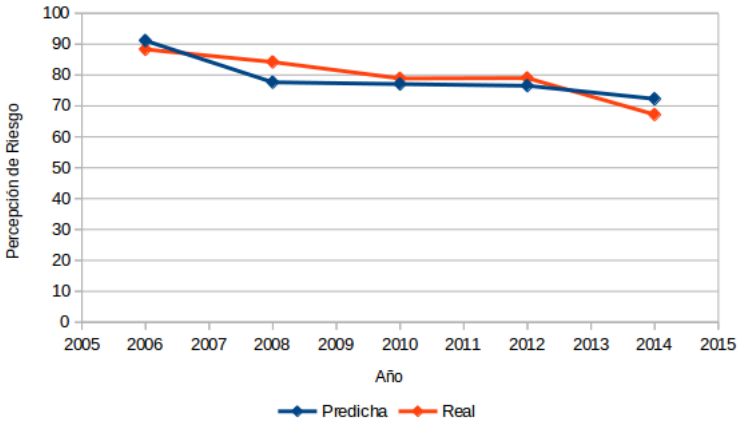

Figura 13: Comparación de Percep. de Riesgo

de percepción de riesgo que 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 13 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.

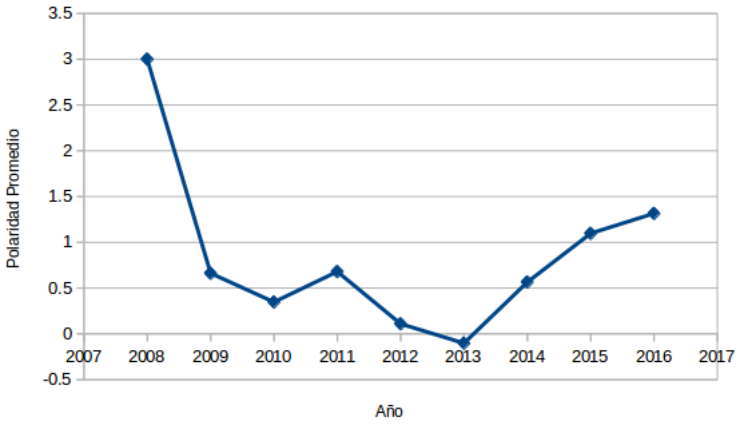

Figura 14: Polaridad en *Tweets* de Políticas

5.6.3. 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 *tweets* clasificados como políticas. La Figura 14 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 de 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.

6. Conclusiones y Trabajo Futuro

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.

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 puedan explicar 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.

Agradecimientos:

Los autores agradecen al Servicio Nacional para la Prevención y Rehabilitación del Consumo de Drogas y Alcohol (SENDA), por su apoyo en la calibración de los modelos desarrollados.

Este trabajo fue parcialmente financiado por el Instituto Sistemas Complejos de Ingeniería ISCI (ICM-FIC: P05-004-F, CONICYT: FB0816).

Referencias

[1] Mir M Ali, Aliaksandr Amialchuk, and Debra S Dwyer. The social contagion effect of marijuana use among adolescents. *PloS one*, 6(1):e16183, 2011.

[2] Jorge Balazs. Diseño, desarrollo e implementación de una aplicación de web oinion mining para identificar el sentimiento de usuarios de twitter con respecto a una compañía de retail, 2015.

[3] Jorge A Balazs and Juan D Velásquez. Opinion mining and information fusion: A survey. *Information Fusion*, 27:95–110, 2016.

[4] Michael Chary, Nicholas Genes, Andrew McKenzie, and Alex F Manini. Leveraging social networks for toxicovigilance. *Journal of Medical Toxicology*, 9(2):184–191, 2013.

[5] Michael J Cleveland, Mark E Feinberg, Daniel E Bontempo, and Mark T Greenberg. The role of risk and protective factors in substance use across adolescence. *Journal of Adolescent Health*, 43(2):157–164, 2008.

[6] Stephanie H Cook, José A Bauermeister, Deborah Gordon-Messer, and Marc A Zimmerman. Online network influences on emerging adults' alcohol and drug use. *Journal of youth and adolescence*, 42(11):1674–1686, 2013.

[7] Sistema 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.

[8] Susan T Ennett, Karl E Bauman, Andrea Hussong, Robert Faris, Vangie A Foshee, Li Cai, and Robert H DuRant. The peer context of adolescent substance use: Findings from social network analysis. *Journal of research on adolescence*, 16(2):159–186, 2006.

[9] Irving J Ginsberg and James R Greenley. Competing theories of marijuana use: A longitudinal study. *Journal of Health and Social Behavior*, pages 22–34, 1978.

[10] Natalia Hernández. Metodología para el diseño y construcción de un lexicón de opinion mining, basado en comentarios de twitter aplicado al proyecto "opinionzoom", 2016.

[11] Katherine M Keyes, John E Schulenberg, Patrick M O'Malley, Lloyd D Johnston, Jerald G Bachman, Guohua Li, and Deborah Hasin. The social norms of birth cohorts and adolescent marijuana use in the united states, 1976–2007. *Addiction*, 106(10):1790–1800, 2011.

[12] Kimberly Kobus and David B Henry. Interplay of network position and peer substance use in early adolescent cigarette, alcohol, and marijuana use. *The Journal of Early Adolescence*, 2009.

[13] Joseph W LaBrie, Justin F Hummer, and Andrew Lac. Comparing injunctive marijuana use norms of salient reference groups among college student marijuana users and nonusers. *Addictive behaviors*, 36(7):717–720, 2011.

[14] Gang Lee, Ronald L Akers, and Marian J Borg. Social learning and structural factors in adolescent substance use. *W. Criminology Rev.*, 5:17, 2004.

[15] Edison Marrese-Taylor, Juan D Velásquez, Felipe Bravo-Marquez, and Yutaka Matsuo. Identifying customer preferences about tourism products using an aspect-based opinion mining approach. *Procedia Computer Science*, 22:182–191, 2013.

[16] John Petraitis, Brian R Flay, and Todd Q Miller. Reviewing theories of adolescent substance use: organizing pieces in the puzzle. *Psychological bulletin*, 117(1):67, 1995.

[17] Sarah A Stoddard, Jose A Bauermeister, Deborah Gordon-Messer, Michelle Johns, and Marc A Zimmerman. Permissive norms and young adults' alcohol and marijuana use: The role of online communities. *Journal of Studies on Alcohol and Drugs*, 73(6):968–975, 2012.

[18] Andrea L Stone, Linda G Becker, Alice M Huber, and Richard F Catalano. Review of risk and protective factors of substance use and problem use in emerging adulthood. *Addictive behaviors*, 37(7):747–775, 2012.

[19] Marianne BM van den Bree and Wallace B Pickworth. Risk factors predicting changes in marijuana involvement in teenagers. *Archives of general psychiatry*, 62(3):311–319, 2005.

[20] Mark J Van Ryzin, Gregory M Fosco, and Thomas J Dishion. Family and peer predictors of substance use from early adolescence to early adulthood: An 11-year prospective analysis. *Addictive behaviors*, 37(12):1314–1324, 2012.

[21] Juan D Velasquez, Alejandro Bassi, Hiroshi Yasuda, and Terumasa Aoki. Mining web data to create online navigation recommendations. In *Data Mining, 2004. ICDM'04. Fourth IEEE International Conference on*, pages 551–554. IEEE, 2004.

[22] Suzanne L Wenzel, Joan S Tucker, Daniela Golinelli, Harold D Green, and Annie Zhou. Personal network correlates of alcohol, cigarette, and marijuana use among homeless youth. *Drug and alcohol dependence*, 112(1):140–149, 2010.