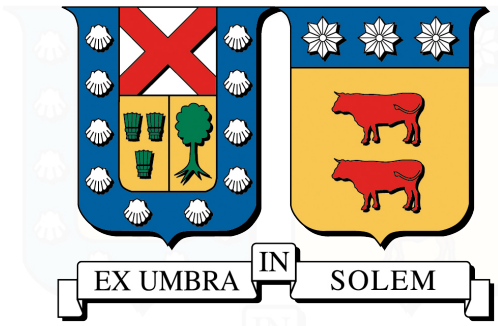
2017

USO DE SNA PARA DILUCIDAR INFLUENCIAS ENTRE EL RECLAMO SOCIAL Y LA PRENSA TRADICIONAL EN TWITTER

SIMONSEN SIMONSEN, AXEL VAN

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INFORMÁTICA
VALPARAÍSO - CHILE



USO DE SNA PARA DILUCIDAR INFLUENCIAS ENTRE EL RECLAMO SOCIAL Y LA PRENSA TRADICIONAL EN TWITTER

AXEL VAN SIMONSEN SIMONSEN

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL INFORMÁTICO

PROFESOR GUÍA	:	SRA. CLAUDIA LOPEZ
PROFESOR CORREFERENTE	:	SR. IGNACIO ARANA

SEPTIEMBRE 2017

EX UMBRA
IN
SOLEM

A mi familia y amigos, forjadores de esta etapa . . .

EX UMBRA
IN
SOLEM

Resumen Ejecutivo

La investigación en *social media* se ha convertido en un tema de gran interés en las ciencias sociales. Sin embargo, la investigación en esta área requiere procesamiento y análisis de potencialmente grandes volúmenes de datos. Esto abre un área de colaboración entre la informática y las ciencias sociales.

Esta memoria presenta una metodología de investigación de datos provenientes de *Twitter* que combina *Social Network Analysis* (SNA), análisis estadístico descriptivo y análisis de sentimientos para examinar la influencia de diferentes actores sociales en la discusión pública. La metodología fue aplicada a un caso de estudio formado por la discusión en Twitter sobre cuatro reformas sociales impulsadas por el gobierno de Michelle Bachelet. Los resultados revelaron el nivel influencia en la discusión pública de cuatro actores clave: el poder ejecutivo, legislativo, los medios de prensa tradicional y la ciudadanía.

Abstract

Research on social media has become an interesting topic to study for social scientists. However, this kind of research requires processing and analyzing potentially large volumes of data. This need opens a new collaboration area between informatics and the social sciences.

This project proposes a methodology to investigate data from Twitter that combines social network analysis, descriptive statistics and sentiment analysis to examine to what extent different social actors influence the public discussion. The methodology was applied to a case study comprising the Twitter discussion on four social reforms undertaken by Michelle Bachelet's government. The results revealed how influential are the legislative and executive power, traditional media and citizens in shaping the public debate on Twitter.

Keywords: Social Media, Social Network Analysis, Sentiment Analysis, Twitter, Graph Theory

EX UMBRA
IN
SOLEM

Índice de Contenidos

Índice de Tablas

EX UMBRA
IN
SOLEM

Índice de Figuras

EX UMBRA
IN
SOLEM

Introducción

En la historia de un país, existen muchos eventos y fenómenos que en su conjunto y cronológicamente, explican la forma de ser, de vivir, y en general, la cultura de sus ciudadanos. Cada uno de estos fenómenos sociales es dependiente del contexto histórico y temporal no solo del país en cuestión, sino que también del contexto de sus países vecinos y hoy en día del contexto global. Un fenómeno social actual en Chile, son las movilizaciones sociales, siendo las mas notorias las movilizaciones estudiantiles que por los últimos 11 años se han tomado las calles creando el *"Movimiento Estudiantil"*, uniéndose a lo largo de los años trabajadores, profesores, entre otros actores llamándose hoy: el *"Movimiento Social"*.

Si bien los movimientos sociales no son algo de estos últimos años, si no que son fenómenos que recurrentemente ocurren en la historia de las civilizaciones, el uso de la tecnología le ha dado nuevos matices al desarrollo de las protestas sociales, permitiendoles evolucionar como fenómeno social. Por otro lado, el uso de las aplicaciones de *Social Media* como Facebook y Twitter permiten a la ciudadanía opinar sobre los temas de contingencia, organizar campañas informativas y eventos tales como marchas y manifestaciones de diversas índoles (García et al., 2013; Cárdenas Neira, 2014). Todos estos fenómenos que unen a la sociedad y el uso de la tecnología son temas de interés para investigar a través de los datos disponibles en Facebook y Twitter el comportamiento social de una red de usuarios, utilizando un análisis llamado *Social Network Analysis*. Sin embargo, no solo los ciudadanos han logrado sacar provecho de estas aplicaciones... actores como los medios de prensa han entrado en este nuevo medio de comunicación, siendo un canal que les permite entregar su contenido al resto de usuarios de manera mas cercana que la tradicional, como periódicos y páginas web (Hong, 2012). De la misma manera los políticos han utilizado estos medios para debatir entre ellos y para comunicarse principalmente con la red de usuarios, con el objetivo de conseguir adherentes y hacer campañas en periodos electorales.

En esta investigación se centra el contexto social en las cuatro grandes reformas chilenas impulsadas por el último gobierno de Michelle Bachelet: La reforma educacional, tributaria y laboral. Se añade una "reforma constitucional", que si bien no es una reforma promulgada como tal, se contextualiza en un proceso de cambio anunciado por el gobierno de Michelle Bachelet.

Analizando lo que los usuarios en Twitter publicaron sobre estas reformas, se pretende conocer cual es el modelo de difusión de la información, encontrando a los usuarios influyentes en opinión sobre estos tópicos y si se replica el modelo "offline", donde los medios de prensa tienen el control de la mayor parte de la información existente. Para lograr esto se realizaron 3 tipos de análisis sobre los datos: Descriptivo, de redes y de sentimientos, el primero para describir el comportamiento de los datos, mientras que los siguientes permiten responder las preguntas de investigación planteadas en esta memoria. Para esto, el documento presenta 6 capítulos. En el primer capítulo se presenta el contexto de desarrollo de este trabajo, junto al problema que motiva esta investigación y los objetivos de la misma. El segundo capítulo muestra como se ha utilizado SNA en otros casos a nivel mundial incluyendo otros casos chilenos. El capítulo tres muestra la metodología de los diferentes análisis, y el framework de trabajo propuesto para realizar el análisis de datos. El cuarto capítulo muestra los resultados de aplicar dicha metodología a los datos recolectados y los análisis a estos resultados. El capítulo cinco muestra la validación de la metodología propuesta para los análisis pertinentes, y finalmente el sexto capítulo muestra las conclusiones del trabajo realizado.

Contexto

Al 2016, la población chilena se estimaba sobre las 18 millones de personas. [1] En este número existe una alta tasa de alfabetización de 97.5 % y altas tasas de acceso a internet (71.5 % a nivel de hogares y 33.1 % en acceso movil), [2] mientras que en internet movil hay una tasa de penetración del 37 %, [3], respaldado por datos de Adimark (GfK Adimark, 2016).

El Internet, se encuentra presente en prácticamente cada momento de nuestras vidas, siendo el término "conectividad" un concepto frecuente y una meta constante por su expansión. Un ejemplo claro de aumento de conectividad fue la inclusión del Internet en los teléfonos celulares (hoy llamados *smartphones*), dispositivos capaces de tomar fotografías, grabar videos y subir este contenido a la red en cuestión de segundos, potencialmente para cualquier usuario en el mundo. Sin embargo, el uso de estas tecnologías se han utilizado para otros propósitos. Es normal hoy en día conocer noticias por canales diferentes a los medios de prensa tradicionales (digase periódicos, televisión y radios), como caza noticias, los cuales son usuarios comunes que suben contenido informativo a las diferentes plataformas existentes, como también organizaciones y medios de prensa no inscritos oficialmente, los cuales funcionan sólo mediante las redes sociales.

[1] http://www.ine.cl/canales/chile_estadistico/familias/demograficas_vitales.php
[2] https://www.cia.gov/library/publications/the-world-factbook/geos/ci.html?
[3] http://www.telefonicachile.cl/wp-content/uploads/2016/12/Informe-Big-Data_20161.pdf

Dado el contexto anterior, los medios tradicionales han tenido que reinventarse, tomándose un espacio en estas redes, siendo Twitter una de las más visibles, permitiendo a las radios y programas de televisión interactuar en tiempo real con su audiencia por medio de esta red (Harrington et al., 2013). En el caso de los periódicos y medios de prensa escritos, permite entregar a sus lectores las noticias de manera inmediata sin tener que esperar a la publicación física. Aunque si bien, esto se lograba mediante los sitios web, el uso de Twitter permite enviar notificaciones a los lectores o simplemente aparecer en la sección de noticias de cada usuario que actúe como seguidor de estos medios de prensa.

Definición del Problema

Dado el contexto actual, donde las plataformas de *Social Media* han pasado a formar parte del estilo de vida de la sociedad, naturalmente, se crea un nuevo flujo de información entre las personas.

Hasta antes de la existencia de estas plataformas, los únicos proveedores de información eran los medios de comunicación tradicionales, los cuales poseían el dominio de la información entregada de acuerdo a las diferentes líneas editoriales de cada uno de estos (McQuail, 1994). Así, la prensa *offline*, funciona bajo el diagrama donde el emisor y receptor están determinados previamente (ver figura 1.1.a). Sin embargo, con el uso de las redes sociales como medios de comunicación el emisor y receptor se convierten en una entidad dinámica que puede cumplir ambos roles (ver figura 1.1.b). Bajo este nuevo modelo, en el cual los medios de comunicación también son usuarios de estas redes, nace el problema de saber si el nuevo modelo replica de alguna forma en el mundo *online* las mismas consecuencias producidas por el modelo *offline*. Es decir, **¿Tienen los medios de comunicación el control de la información circulante en las redes sociales?, ¿Se replica el modelo offline en el nuevo contexto online?**.

Si bien es normal suponer que en la red existen usuarios mas influyentes que otros, si se replica el modelo que existen con los medios de prensa fuera de la Internet, la información en redes sociales no está aportando un valor agregado como medio alternativo de comunicación. Sin embargo, no es fácil reconocer a simple vista como viaja la información ni quienes son las personalidades influyentes en estas redes, siendo este, el primer problema elegido a ser resuelto, mediante las siguientes preguntas de investigación.

- ¿Quienes son los usuarios influyentes en una red de usuarios en un contexto determinado?
- ¿Están las redes sociales realmente dando libertad y visibilidad de las opiniones de cada usuario por igual?
- ¿Existe un dominio de la prensa en la información online?
- ¿Existe dominio por parte de ciertos actores sociales?
- ¿Que clase de información es la que se difunde a través de la red?

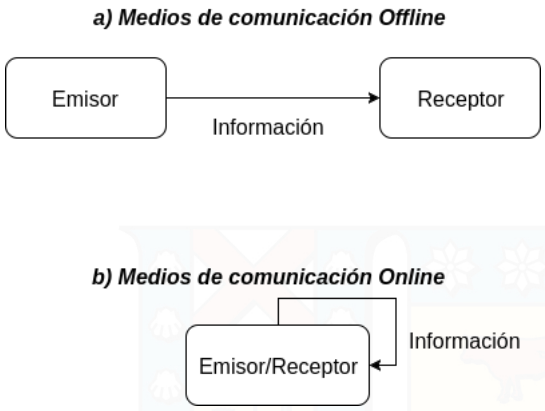


Figura 1.1: Flujos de información offline y online

(Fuente: Elaboración Propia)

Sin embargo, el principal problema es lograr **realizar un análisis confiable, de manera automatizada, que sea capaz de responder las preguntas de los investigadores, con un volumen de datos, el cual es imposible de analizar sin herramientas computacionales**. Este problema debe ser dividido en sub problemas, los cuales son:

1. Como recolectar los datos de manera que éstos posean relevancia con los temas sociales estudiados, eligiendo una plataforma de *Social Media* adecuada.
2. Identificar los actores sociales sobre los que se quiere averiguar su comportamiento.
3. Determinar métricas sobre como medir cuando un usuario es influyente en la red.
4. Determinar como medir la equidad de las redes en un conjunto de usuarios.

Propuesta de solución

Ante el problema existente, se propone como solución el desarrollo de un framework de análisis de datos mediante *Social Network Analysis* utilizando Twitter como plataforma proveedora de los datos. Ésta elección se debe al uso que le dan los medios de comunicación en Chile a esta plataforma y la capacidad de catalogar la data mediante *hashtags*, los cuales son etiquetas para catalogar los tweets, creadas por los mismos usuarios, permitiendo buscar por tópicos, sobre los que los usuarios publican.

Con la plataforma social escogida, se ha elegido tocar un tema de índole político, dado que posee la cobertura de los medios de prensa y la ciudadanía, ambos publicando en Twitter. Por otro lado, se incluirán actores de carácter político, como son los diputados y senadores. El objetivo es comprobar la influencia de cada actor en el tópico investigado, el cual dada la contingencia política existente en el momento de comenzar este proyecto, será el proceso de reformas del gobierno de Michelle Bachelet, entre las que se encuentran la reforma Educacional, Laboral, Tributaria y Constitucional. Ésta última si bien no es una reforma en si misma, está en el proceso de cabildos ciudadanos, con el objetivo de redactar una nueva constitución.

Éstos datos extraídos vienen en formato JSON, lo que permite un acceso rapido a cada uno de sus campos, sin embargo, aún así se encuentran en bruto, es decir, vienen tal y como los entrega la API de twitter, por lo que para hacer más fácil el análisis, serán procesados para obtener datasets en forma de dataframes, las cuales son estructuras que agregan metadata sobre sus filas y columnas a los datos, lo cual permite aplicar consultas SQL sobre estos, facilitando la lectura de los registros, por sobre el acceso directo a los archivos JSON, permitiendo añadir nuevas columnas que permitan recabar mas información de los datos.

Con esto, los datos serán sometidos a un análisis descriptivo, SNA (*Social Network Analysis*) y análisis de sentimientos . El primero de estos se realizará con el objetivo de conocer el comportamiento de los datos en cuanto a los hashtags seleccionados como filtro de los datos y en cuanto a los distintos actores participantes del debate en Twitter. SNA será el núcleo de esta memoria, creando las redes de usuarios participantes y determinando quienes son los personajes influyentes, dato que nos permitirá responder la pregunta central de investigación y comprobar si se replica o no el modelo de influencia de la prensa tradicional. Como se puede ver en la figura 1.2, se utilizará la plataforma Twitter para analizar data relacionada a las reformas anteriormente mencionadas. Por último, el análisis de sentimientos se realizará para complementar los análisis anteriores y conocer a través del comportamiento de los datos, cual es la reacción general de la gente mediante las publicaciones que realizan sobre estos proyectos políticos, y como influyen los sentimientos positivos o negativos sobre la red.

Para el desarrollo de esta metodología de investigación se deben utilizar tecnologías que permitan cumplir los siguientes objetivos, según el proceso definido en la figura 1.2:

- Recopilación de los datos mediante Streaming
- Análisis estadístico de los datos
- Representación visual de los resultados
- Procesamiento de grafos
- Creación de modelos y clasificadores

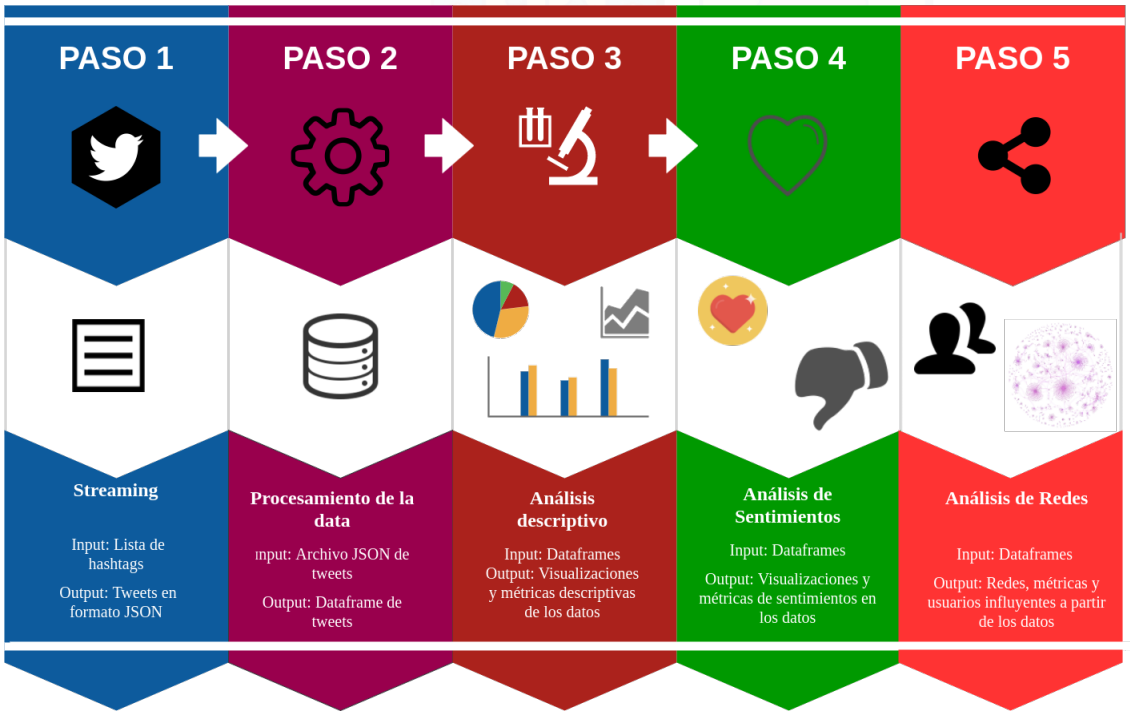


Figura 1.2: Propuesta de solución para análisis de datos

(Fuente: Elaboración Propia)

Objetivos

Dada la solución propuesta, esta debe ser capaz de responder a la problemática definida anteriormente, mediante métricas que permitan responder las preguntas de investigación planteadas y utilizando un marco teórico que permita darle confiabilidad a las respuestas encontradas. Para validar esto, se plantean los siguientes objetivos a ser cumplidos por esta memoria.

Objetivo General

1. Combinar técnicas de recolección y análisis de datos automatizada que permita evaluar si Twitter, como medio de comunicación social y abierto, conserva el modelo de influencia unidireccional desde la prensa a la ciudadanía, o ha dado paso a cambiar dicho modelo.

Objetivos Específicos

1. Definir un método de recolección de datos relevantes a la agenda pública y al tema social escogido.
2. Proponer una metodología de análisis de influencias de distintos actores sociales en temas de interés público en Twitter
3. Validar la metodología creada aplicada al caso de las cuatro grandes reformas chilenas del gobierno de Michelle Bachelet.

EX UMBRA
IN
SOLEM

Estado del arte

Este capítulo presenta el estudio y como surge una de las ramas de la ciencia que une los mundos de las ciencias sociales y las ciencias computacionales, a través del estudio de aplicaciones informáticas conocidas como *Social Media* y la utilización de distintos algoritmos diseñados para conocer más de la sociedad utilizando los datos de estas aplicaciones. Esta memoria se sitúa en esta rama debido a que busca conocer sobre la agenda pública y la influencia entre los usuarios chilenos en Twitter. Por último se presenta una revisión breve a la literatura de *Social Network Analysis* y casos de aplicación de esta rama de investigación.

La informática en estudios sociales

Pasados los años 50, como una mezcla entre disciplinas tales como filosofía, bibliotecología, lingüística, matemáticas, ciencias sociales, entre otras, nace el campo de estudio denominado como *Information Sciences* (Wersig y Neveling, 1975). Con el nacimiento de tecnologías de *Social Media*, esta disciplina comienza a trabajar con estas tecnologías, dado que poseen grandes cantidades de datos a nivel global. En el caso de Twitter, donde se emiten alrededor de 7000 tweets por segundo[4], se trata de una gran cantidad de datos para analizar sobre distintas temáticas. Esta alza masiva en el volumen de datos para investigar generó una oportunidad para realizar estudios de la sociedad misma en la dimensión online, complementando estos datos provenientes de las personas mismas con estadísticas gubernamentales o información recabada directamente a un determinado número de personas utilizando los métodos clásicos, como por ejemplo, las encuestas.

[4]http://www.internetlivestats.com/one-second/#tweets-band

Esta rama de investigación, con más de 20 años de vida, hoy en día trabaja con una gran cantidad de información como fuentes de datos para investigación, gracias a medios como Facebook y Twitter. Estos sistemas poseen cada uno una API (Application Programming Interface), las cuales son sistemas de interacción entre medios externos y la aplicación, lo que permite realizar recolección de datos para desarrolladores externos. Esto las convierte en una potencial fuente de datos para responder distintas preguntas de índole social, utilizando diferentes técnicas y metodologías, como por ejemplo, el análisis de grafos.

Twitter como plataforma de investigación

Twitter es una red social de distribución de contenido del tipo *microblogging*. Este servicio permite que distintos usuarios puedan comentar cosas tanto para sus seguidores, como también escribirle a personas no conocidas por el usuario. Esta característica es una de las que mas valor da a Twitter dado que, en teoría, se eliminan las barreras que impiden la comunicación con ciertas personas dados sus cargos, como políticos, personajes famosos, etc.

Twitter comenzó el año 2006, fundada por Jack Dorsey, Biz Stone y Evan Williams y al año 2010 poseía 30 millones de usuarios. A finales del 2016, la aplicación ya ha llegado a tener 319 millones de usuarios, como lo revela la figura 2.1 y los datos publicados en su sitio web con información sobre la compañía[5].

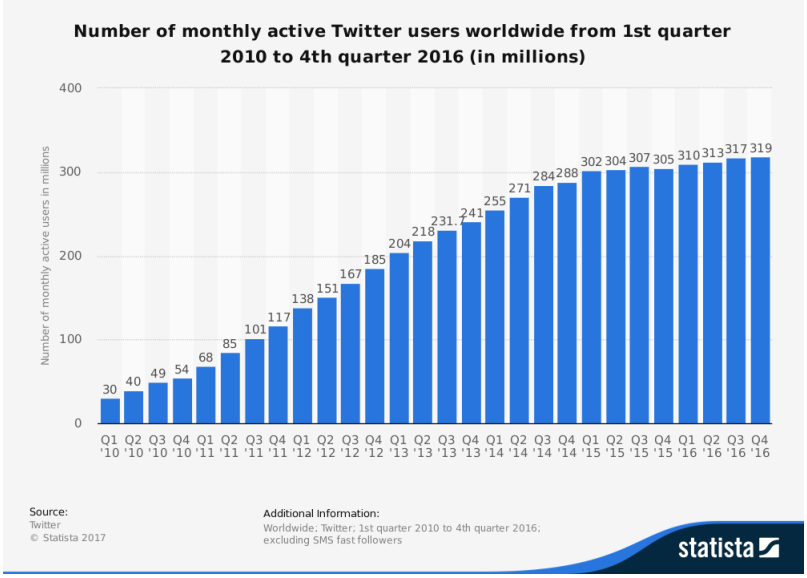


Figura 2.1: Gráfico de usuarios de Twitter en el tiempo

(Fuente: Statista, www.statista.com)

[5]https://about.twitter.com/es/company

En Chile, según el libro "Encuesta Nacional Bicentenario" (GfK Adimark, 2016), el 56 % de las personas encuestadas declara conocer Twitter, siendo el perfil más común los usuarios entre los 18 y 34 años del estrato alto de la sociedad (ver figuras 2.2, 2.3 y 2.4).

Se escoge Twitter como ambiente de estudio, dado que los distintos actores a estudiar poseen cuentas oficiales, esto se complementa muy bien con la característica de Twitter de permitir el debate entre personas que no necesariamente se conocen, en distintos lugares del mundo y sobre cualquier tema, permitiendo conversar con políticos y medios de comunicación, como declaran los usuarios encuestados en la tabla 2.1, donde el 56 % de los usuarios de Twitter encuestados declara seguir a medios de comunicación, y el 21 % declara seguir las cuentas de los políticos.

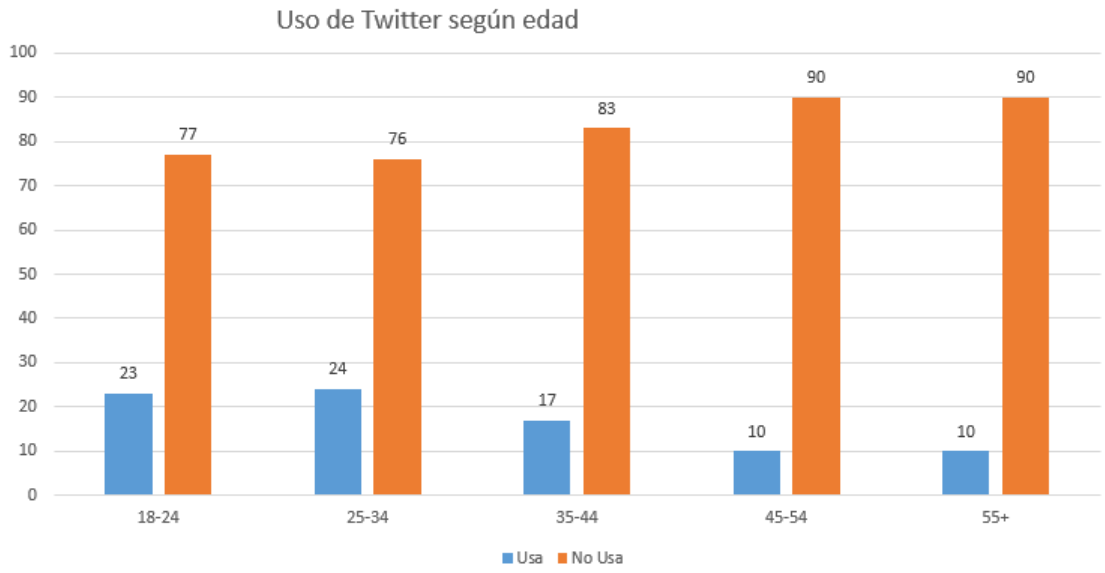


Figura 2.2: Encuesta Adimark: Uso de Twitter según edad

(Fuente: Encuesta Nacional Bicentenario, Adimark)

Hecho	Porcentaje del total de usuarios de Twitter
Seguidores de Medios de comunicación	56 %
Seguidores de Políticos	21 %

Tabla 2.1: Estadísticas Adimark de usuarios de Twitter en Chile

(Fuente: Encuesta Nacional Bicentenario, Adimark)

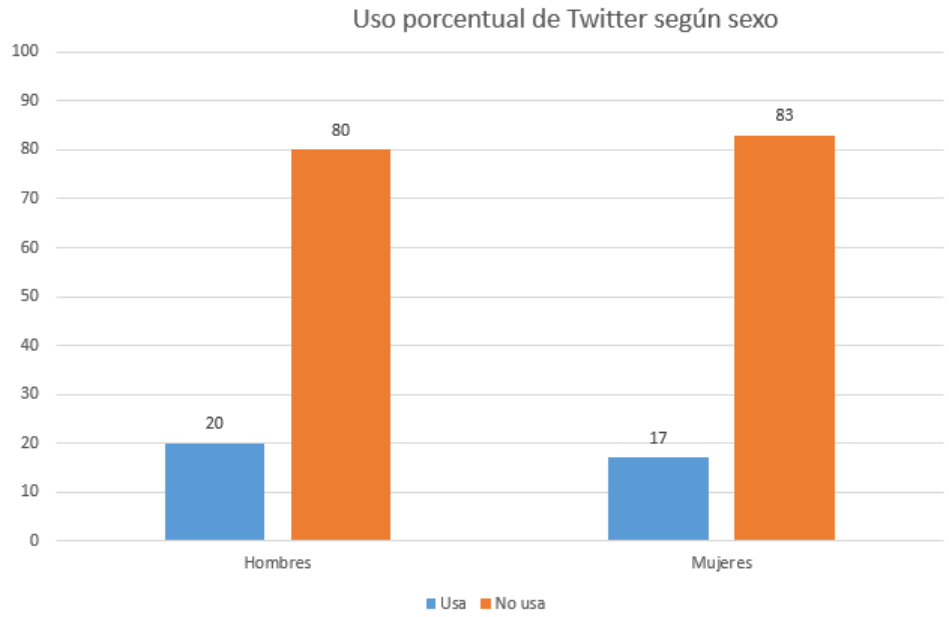


Figura 2.3: Encuesta Adimark: Uso de Twitter según sexo

(Fuente: Encuesta Nacional Bicentenario, Adimark)

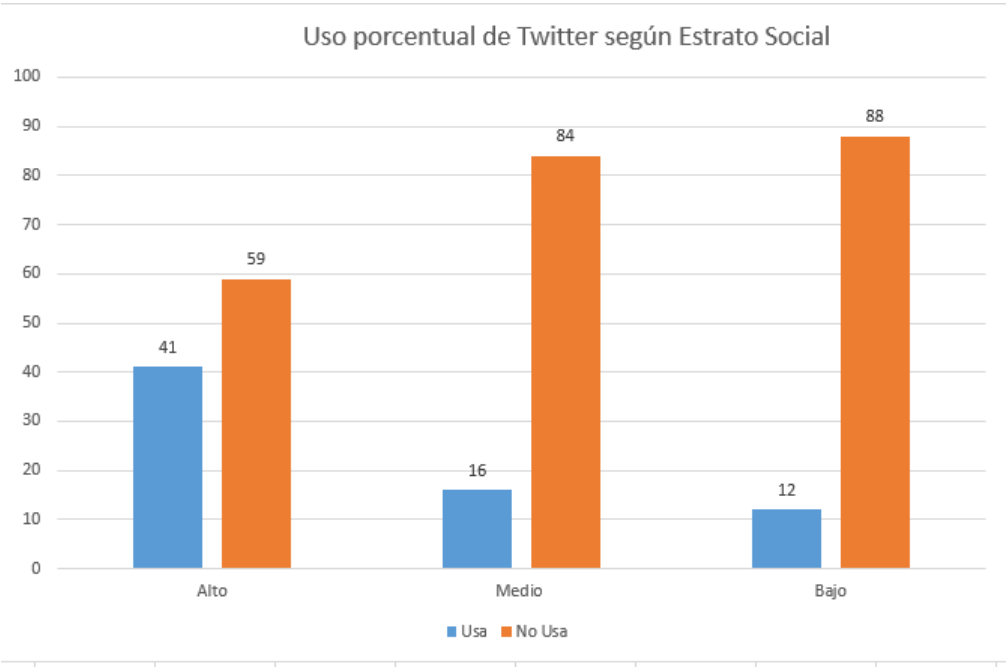


Figura 2.4: Encuesta Adimark: Uso de Twitter según estrato social

(Fuente: Encuesta Nacional Bicentenario, Adimark)

Revisión de la literatura: Twitter y la Agenda Pública

Uno de los principales puntos a investigar trata sobre influencia en la agenda pública que, basado en la teoría de Configuración de la Agenda (McCombs y Shaw, 1972), los temas a los cuales los medios dan cobertura condicionan lo que la ciudadanía discute y otorga importancia en un momento determinado. Esta teoría, validada en su tiempo, es discutida hoy respecto a su aplicabilidad en la realidad actual, dada la cantidad de medios de comunicación que la gente puede elegir (Bennett y Iyengar, 2010), siendo esta teoría aplicada en un entorno donde existía un monopolio de la información transmitida a la ciudadanía. De la misma forma también se discute su vigencia, añadiendo factores ligados a la predisposición de las personas a escoger entre distintos medios y canales de comunicación, como el escrito, el visual, entre algunos otros, complementando la teoría (Scheufele y Tewksbury, 2007).

Gracias a las redes sociales, la variedad de fuentes de información son muchísimas (Meraz, 2009a), sin embargo, la credibilidad de estas, no es siempre la misma. En esta línea, esta investigación pretende conocer un poco más sobre como la llegada de las redes sociales (en este caso Twitter) ha influido en como es influenciada la agenda pública.

Otro punto importante es la teoría sobre la existencia de una élite en los usuarios de Twitter. Esta élite, manejaría en su mayoría lo que se discute en la red, como usuarios influyentes de la misma sobre el resto de usuarios normales (Tumasjan et al., 2010; Wu et al., 2011; Kulshrestha et al., 2012).
Según (Wu et al., 2011), el 0,05 % de los usuarios de Twitter habían publicado cerca de la mitad del contenido total. Esto se podría complementar con la configuración de la agenda pública, dado que si existen élites o minorias que son la fuente del contenido en Twitter, se replicaría la situación “offline” donde las lineas editoriales controlan el contenido que se publica. Una técnica utilizada para analizar el comportamiento de redes sociales y su flujo de información es *Social Network Analysis*, la cual permitiría comprobar si efectivamente este es el modelo bajo el cual funciona la red de Twitter en Chile.

Técnicas computacionales utilizadas en Twitter

Social Network Analysis

El área de *Social Network Analysis* ha ido evolucionando desde la década de los ochenta hasta el día de hoy, gracias a la evolución de la tecnología a nivel de algoritmos, procesamiento de datos y almacenamiento de los mismos. Uno de los libros insignes de esta disciplina (Wasserman y Faust, 1994) aplica gran parte de la teoría de grafos y modelos estocásticos al análisis de grafos, representando relaciones sociales en un contexto offline dado que en la época no existía ni el internet, ni mucho menos las redes sociales. Para estos casos, los datos podían ser relaciones entre personas y/o organizaciones, mientras que los arcos relacionales podían representar relaciones económicas, afectivas, políticas, entre muchas otras, ya que básicamente cualquier interacción social podía ser representada gracias a estas redes permitiendo responder nuevas preguntas sobre el comportamiento social que antaño no poseían respuestas, como por ejemplo el fenómeno de agrupación de determinados grupos de personas en relación a ciertos patrones y la afiliación entre organizaciones y personas.

Posterior a esto, en el año 2005, Wasserman nuevamente llevaba un paso adelante la disciplina, añadiendo a estas redes el calculo por medio de modelos y métodos matemáticos (Carrington et al., 2005), aprovechándose también de la representación matricial que las redes poseen. Esta fue la puerta de entrada a un análisis de redes sociales que permitiría en un futuro aplicar algoritmos cada vez mas complejos mas allá del análisis de grafos, lo cual sería de importancia vital a la disciplina en los años venideros.

A la llegada de las aplicaciones de *Social Media*, principalmente Facebook y Twitter junto a los avances en la algorítmica permitió analizar estos datos con técnicas como *Content Analysis, Machine Learning*, entre otras. Así, el interés ya no solo existía en las relaciones entre las personas (algo aún muy estudiado), sino que se agrega también, el análisis del contenido de las publicaciones, siendo una de las mas utilizadas, el Análisis de Sentimientos (Kouloumpis et al., 2011).

Sentiment Analysis

El análisis de sentimientos es una de las técnicas de procesamiento de lenguaje natural (*Natural Language Processing* o NLP), que consiste en clasificar y reconocer la emoción en un texto. Esta puede ser una emoción en específico o reconocer la tendencia de la emoción (positiva, negativa o neutra), a través del uso de algoritmos relacionados al aprendizaje de máquinas.

Este proceso puede ser hecho a nivel de frases o a nivel de documentos completos (Feldman, 2013). Ambos procesos, han sido probados utilizando distintas técnicas de aprendizaje, debido a los problemas existentes al momento de analizar lenguajes, siendo la ambigüedad uno de los más grandes.

La teoría de lenguajes formales explica que un lenguaje puede tener ambigüedades. Los lenguajes de programación por ejemplo, deben ser creados para que no existan ambigüedades, lo cual se resuelve con el uso de paréntesis, expresiones lógicas, entre otros símbolos y propiedades del lenguaje. Sin embargo, los lenguajes utilizados por las personas, no poseen todas estas reglas para evitar ambigüedades, por lo que dependiendo del contexto pueden significar dos o más cosas muy diferentes. Ante este problema, los algoritmos de *Machine Learning* son muy afectados, reduciendo sus valores de precisión. Estos problemas han logrado ser reducidos a través de modificaciones en los algoritmos mismos, estimadores de significados de frases según contexto (Roth, 1998), entre otras técnicas que se han desarrollado (Wilson et al., 2005).

Una vez mitigado el problema, algunas de las técnicas mas confiables utilizadas han sido las *Support Vector Machines* (SVM) para aprendizaje supervisado (Mullen y Collier, 2004). También se han creado modelos utilizando Redes neuronales (Dos Santos y Gatti, 2014), y otros mecanismos de aprendizaje no supervisado, utilizando modificadores para los casos de uso de los llamados “emoticones”, los cuales al poseer una ponderación propia en los modelos pueden mejorar la clasificación (Hu et al., 2013).

Una de las mayores aplicaciones de esta disciplina es conocer la opinión general de un conjunto de personas respecto a ciertos tópicos. Es aquí cuando se trabaja con datasets contextuales para tópicos políticos, deportivos, relacionados al cine, la música, etc. Esto se complementa de excelente manera con la llegada de Twitter (Pak y Paroubek, 2010). Su sistema de *microblogging*, permitía tener diferentes opiniones, en textos pequeños de distintas personas. Además, el sistema de *Hashtags* permitía separar estas opiniones según tópicos determinados. Es aquí cuando se comienza a utilizar la data de Twitter y otras aplicaciones de *Social Media* para Procesamiento de Lenguaje Natural, sumándose como una herramienta poderosa complementaria al uso de SNA.

Actualmente el idioma inglés es el idioma estándar a nivel mundial para la comunicación de negocios, tecnologías y relaciones internacionales, por lo tanto la mayoría del estado del arte existente en Análisis de Sentimientos está desarrollada con datos en idioma inglés. Distinto es el caso del idioma español, donde el procesamiento de lenguaje natural en este idioma se encuentra en una etapa mucho mas inmadura.

Actualmente, y ya como disciplina en pleno auge, *Social Network Analysis* encuentra el problema del procesamiento de grandes cantidades de datos, dada la masividad de usuarios, contenido disponible y complejidad de las técnicas algorítmicas. A partir de este problema, la disciplina se ha unido con el área de big data utilizando procesamiento distribuido, optimizando los tiempos de procesamiento mediante concurrencia (Ediger et al., 2010), permitiendo analizar conjuntos mucho más grandes de datos.

SNA: La prensa y los medios digitales

Estudios anteriores han investigado la influencia de la prensa utilizando dos medios en internet: las aplicaciones de blogs (Meraz, 2009b) y las plataformas de *Social Media*, buscando encontrar relación entre los cambios en la agenda pública y las publicaciones en dichos blogs como una forma de comprobar la restauración de la equidad en la información. En esta investigación, se llegó a la conclusión que existía un avance respecto a la equidad en los sitios denominados como *Citizen Media*, sitios donde ciudadanos particulares y no necesariamente periodistas, publicaban sus noticias. Algunos de estos sitios se acercaban a los sitios oficiales de la prensa tradicional en Estados Unidos, en cuanto a alcance de usuarios. Estos resultados ya hablaban sobre un cambio en la distribución de la información gracias al internet.

Posterior al nacimiento y ascenso en el uso de Twitter,se intenta investigar la relación entre los tópicos que se informan en el New York Times en relación a lo que hablan los usuarios en Twitter en la misma región geográfica (Zhao et al., 2011). Esta investigación descubrió que si bien existía una relación en los temas hablados, existían tópicos mas debatidos en Twitter respecto a la prensa tradicional, encontrando que además de existir relaciones entre ambos medios, también Twitter posee cierta independencia en su creación de información.

Twitter y movimientos sociales

Cercano al 2011, se levantaba un movimiento social en España, llamado **Los Indignados**. Este movimiento en el área de *Social Media* fue muy particular debido a que su organización fue facilitada gracias a redes sociales utilizadas para difundir las movilizaciones y convocatorias sociales varias (Vallina-Rodriguez et al., 2012). Se analizaron mediante SNA alrededor de 3 millones de tweets de 500.000 usuarios, descubriendo que en comparación a los actores en la política española que poseían cuentas en Twitter, la influencia de la gente y de colectivos sociales es mucho mayor en esta red, dando datos para predecir que este comportamiento social aumentará y las revoluciones serán “twitteadas”.

Para el caso de la **primavera Arabe** en 2011 (Huang, 2011), destaca la importancia de Facebook y Twitter en el alzamiento y la convocatoria de la ciudadanía, con datos donde la mayoría de los países como Egipto, Tunisia, entre otros, vio incrementada su cantidad de usuarios en estas redes entre un 15 y 30 % en tan solo un año. Mientras que la mitad de la población estudiada en este artículo declara que estas aplicaciones fueron una ayuda importante para la movilización.

El caso de Chile

Chile, gracias a sus constantes manifestaciones estudiantiles, las cuales han sido más notorias desde el 2006 gracias a la revolución pingüina, ha estado en la mira de muchos investigadores estudiando la conducta de estas revoluciones, en conjunto con su uso de la tecnología, específicamente, *Social Media*. En las protestas del 2011, las cuales duraron aproximadamente 8 meses, se realizó una investigación para relacionar el uso de Facebook con la organización de este movimiento social (Valenzuela et al., 2012). En esta se analiza Facebook según su uso como medio de información y su uso para la organización de las manifestaciones sociales.

Las conclusiones obtenidas fueron muy similares a las estudiadas en la primavera árabe, es decir, que *Social Media* es un medio eficaz para la organización comunitaria y social.

A diferencia de la investigación anterior, la cual obtenía sus datos de distintas fuentes como Facebook, sujetos encuestados, etc, esta investigación obtuvo sus datos desde Twitter (García et al., 2013), sin necesidad de otras fuentes de datos. Con lo anterior, y realizando un estudio y conexión con los sucesos de la revolución pingüina en 2006, se define una división entre el mundo "online" y el mundo "offline", donde el primero aporta a las actividades de carácter de protesta realizadas en el mundo offline, mientras que los sucesos ocurridos en el segundo afectan y condicionan en cierto grado lo que se habla en el mundo online.

Por otro lado, utilizando SNA, se estudió la evolución de las redes en estos años llegando a la conclusión que entre el 2011 y 2012 la red escaló exponencialmente en su cantidad de usuarios participando en "la revolución online", creciendo también por lo tanto, las redes compuestas por estos usuarios, lo cual demuestra una relación entre *Social Media* y la organización e información de la sociedad.

Conclusiones generales

Como en algunos de los casos anteriores, existen muchísimos más casos aplicados, donde SNA ha jugado un rol importante en descubrir comportamientos de la sociedad en las aplicaciones de *Social Media*. Si bien en sus comienzos, la tecnología solo permitía el uso de blogs y sitios mas "estáticos", ya en esta etapa se podía ver un cambio en la forma de informarse, donde los usuarios comenzaban a crear una independencia a los medios tradicionales, dada la aparición de otras fuentes de información.

En adición a lo anterior, en cuanto a temas políticos, se puede concluir que el uso de *Social Media* no solamente es un medio de comunicación dinámico en cuanto a la información, sino que se ha vuelto una herramienta de organización y discusión para los ciudadanos que les permite manifestarse, a diferencia de los medios tradicionales. Todo esto ha situado al uso de SNA como una herramienta sumamente importante para entender desde una arista diferente, algunos fenómenos sociales que no podrían ser plenamente descritos mediante la investigación de solamente el mundo "offline".

EX UMBRA
IN
SOLEM

Metodología aplicada en la solución

Para lograr obtener respuesta a las preguntas de investigación planteadas en esta memoria, no solo se utilizará *Social Network Analysis* como mecanismo de análisis. Se utilizará también Análisis de sentimientos para poseer un mejor conocimiento sobre la data analizada. Este capítulo comenzará describiendo dichas técnicas (SNA y Análisis de Sentimientos). Posteriormente se describirá en detalle la metodología utilizada, esta se dividirá en tres puntos críticos: Selección de la muestra, recolección de datos y proceso de análisis.

Análisis de grafos, Social Network Analysis y Social Media

El análisis de grafos o de redes usa la teoría de grafos para explicar de manera matemática el comportamiento de una red compuesta de arcos y vértices. Sin embargo, su aplicación en casos reales nace a través de la representación de ciertas situaciones, como redes electrónicas, moleculares, neuronales, redes de sistemas computacionales, entre otras. Estas redes podían ser modeladas mediante un grafo donde sus nodos y arcos representaban diferentes cosas dependiendo del contexto, pero con el patrón común de representar algún tipo de relación representada por los arcos entre las entidades involucradas representadas por los vértices.

Además de los contextos anteriormente mencionados, las redes también son capaces de modelar relaciones sociales. Antes de la existencia de *Social Media*, las relaciones sociales "offline" podían ser representadas mediante estos análisis, como por ejemplo las redes de afiliación entre personas y organizaciones. La idea básicamente era explicar ciertos fenómenos sociales utilizando las propiedades de la teoría de grafos. Esta disciplina se amplía con la llegada de las redes sociales como Twitter o Facebook permitiendo aplicar estos análisis a un conjunto mayor de datos.

Dado que SNA trabaja sobre redes, la forma en la que estas son construidas es determinante al momento del análisis. Cada relación entre las personas debe ser diseñada correctamente de tal forma que la red sea representativa del entorno social analizado. En este caso, cada vértice (*vertex*) es un usuario y cada arco (*edge*) es la relación entre estos usuarios, comúnmente llamados amigos o seguidores. La principal diferencia entre estos dos conceptos es la mutualidad de la relación 3.1, como se explica a continuación:

- Relación de Amistad: Bidireccional, quiere decir que cada uno puede ver el contenido e interactuar con el otro con los mismos permisos.
- Relación de Seguidor: Unidireccional, lo cual quiere decir que si el usuario A sigue a un usuario B, no implica que el usuario B también pueda interactuar o ver el contenido del usuario A.

Figura 3.1: Relaciones de Amistad y Seguidor en Redes sociales

(Fuente: Elaboración Propia)

En esta memoria la fuente de datos será Twitter. Esta plataforma de *Social Media* funciona mediante la relación de seguidores, la cual es unidireccional, por lo tanto, para crear redes representativas de este modelo, se deben crear como grafos dirigidos. El objetivo es conocer la interacción entre personas a través de los retweets, y qué personas son la fuente de la información retweeteada.

Para analizar las redes es importante recurrir a la teoría de grafos para dar explicación o conceptualizar ciertos comportamientos. Los conceptos pertenecientes a SNA utilizados en esta investigación son los siguientes:

- Grado: El grado de un vértice es la cantidad de arcos que salen de él, en este contexto representa la cantidad de personas con las que un usuario se relaciona en la red. Para términos generales se habla del grado promedio de una red, lo que explica la cantidad promedio de personas con la que se relaciona cada usuario en la red, o en otras palabras, la media de los grados de cada vértice. En las redes dirigidas el grado de un nodo se separa en su grado de entrada y grado de salida, dependiendo hacía donde apunte el arco.
- Densidad: La densidad de un grafo habla de la cantidad de arcos existentes respecto al máximo número posible de estos. Esto quiere decir que a mayor cantidad de arcos entre los vértices, mayor es su densidad, siendo la máxima, cuando es un grafo completo. Un grafo completo es aquel donde cada vértice se comunica mediante arcos con cada uno de los demás vértices. Cuando un grafo posee una baja densidad, se considera un *sparse graph*. Al contrario, cuando un grafo posee una densidad alta se dice que es un *dense graph*. Dado que se utilizaran grafos dirigidos, la siguiente ecuación es la utilizada para obtener la densidad de la red, donde:
  - $| E |$: Cardinalidad de los arcos
  - $| V |$: Cardinalidad de los vertices.

$$D = \frac{2 \mid E \mid}{\mid V \mid (\mid V \mid -1)} \tag{3.1}$$

- Coeficiente de Agrupamiento (*Clustering Coefficient*): El coeficiente de agrupamiento de un grafo es una medida de agrupación o interconexión con sus vecinos que puede tomar valores entre 0 y 1. A mayor valor de este coeficiente, mayor interconexión existe entre los usuarios que componen la red. En *SNA*, una red con un valor pequeño de este coeficiente se le conoce como "red de mundo pequeño". Una red de mundo pequeño es aquella donde la mayoría de los vértices no son vecinos entre si (Watts y Strogatz, 1998). Llevado al mundo de las redes sociales, esto dice que si bien una red puede estar compuesta entre muchas personas, cada persona conoce a la menor cantidad posible de entre las demás. En pocas palabras, poca interacción social entre los componentes de la red. Este coeficiente es definido en el libro "Collective dynamics of 'small-world' networks" (Watts y Strogatz, 1998), en el que definen el concepto de red de mundo pequeño y este coeficiente como medida de referencia. El coeficiente de agrupamiento para grafos dirigidos se define de la siguiente manera, con:
  - $|\{e_{jk}\}|$: Cantidad de arcos entre los vecinos del vertice i
  - k_i : grado del vértice i
  - E : Conjunto de arcos del grafo

$$C = \frac{\sum_{i=1}^{n} \frac{|\{e_{jk}\}|}{k_i(k_i-1)}}{n}, e_{jk} \in E \tag{3.2}$$

- Conjuntos Independientes: En teoría de grafos, un conjunto independiente (Tarjan y Trojanowski, 1977) es un conjunto de vértices que no son adyacentes entre ellos, siendo un conjunto independiente maximal el conjunto con la mayor cantidad de vértices que cumple la propiedad de ser independiente. La cantidad de vértices en estos conjuntos no siempre es la misma, dependiendo del vértice donde se comience a recorrer el grafo.
- Ponderación (Peso) de los arcos: Para las redes construidas se ponderaron los arcos con un valor que representa el número de interacciones entre ambos en la dirección correspondiente que apunte cada arco.

Análisis de sentimientos y SVM

El análisis de sentimientos, utiliza el procesamiento de lenguajes naturales para clasificar frases o extractos de texto en emociones positivas, negativas o neutras. Para esto, se utilizan algoritmos de *Machine Learning*, específicamente clasificadores, como es el caso de las redes neuronales, SVM (*Support Vectorial Machine*) y regresiones. Para decidir cual método utilizar, se debe considerar la poca madurez del analisis de sentimientos en español, por lo que es necesario considerar un clasificador conocido y confiable. La SVM, es un clasificador bastante estudiado (Mullen y Collier, 2004), el cual es confiable en cuanto a su precisión, siendo para éste caso la mejor opción para experimentar en la clasificación de un idioma nuevo.

Las máquinas de soporte Vectorial (o SVM, Support Vector Machine) son modelos utilizados comúnmente para problemas de clasificación. Pertenecen a la categoría de aprendizaje supervisado, esto quiere decir, que el modelo se crea/entrena a partir de datos para los cuales se conoce el resultado deseado. Generalmente, a mayor cantidad de datos, mejores resultados. Luego se clasifican nuevos datos para los cuales la máquina no conoce su resultado. Teóricamente, una SVM crea un conjunto n-dimensional de hiperplanos los cuales son el límite entre una categoría y otra. Para el caso de dos dimensiones, la SVM correspondiente crea una recta en el plano, la cual divide las categorías. Cuando el modelo es creado y la nueva data es predicha, la clasificación correspondiente hará caer el resultado en algún espacio del plano según dicha categoría. Un punto importante, es que uno de los objetivos al crear una SVM es maximizar la distancia al punto más cercano al hiperplano, es decir, crear un *gap* de separación lo mas amplio posible. Sin embargo, todo lo anterior depende si la data es linealmente separable. Para explicar este concepto, tomaremos como referencia el caso en dos dimensiones, de acuerdo a (Cristianini y Shawe-Taylor, 2000).

1. Data Linealmente Separable: Si la data es linealmente separable, esta puede ser separada por dos planos paralelos, buscando maximizar la distancia entre ambos planos. Estos hiperplanos se calculan de la siguiente forma

$$\vec{w} \cdot \vec{x} - b = \pm 1 \tag{3.3}$$

   Para este caso base, dimensionalmente hablando, encontrar los planos no es más que encontrar la recta correspondiente. Sin embargo, esta metodología es escalable para n dimensiones

2. Data no separable linealmente: Cuando la data no puede ser separada linealmente es usada una función llamada *Hinge Loss*

$$\text{máx}(0, 1 - y_i(\vec{w} \cdot \vec{x} - b)) \tag{3.4}$$

Esta función vuelve cero los puntos que están en el lado correcto del plano, y los que no los vuelve proporcional a la distancia al plano central. Finalmente el objetivo se vuelve minimizar una función sopesando la máxima distancia entre los planos, asegurándose que los vectores $\vec{x_i}$ caigan en el lado correcto del plano.

$$\text{mín}[\frac{1}{n}\sum_{i=1}^{n}\text{máx}(0, 1 - y_i(\vec{w} \cdot \vec{x} - b))] + \lambda||\vec{w}||^2 \tag{3.5}$$

Metodología de Trabajo

Esta sección define el proceso de desarrollo de esta memoria en tres pasos fundamentales: Selección de la muestra, recolección de datos y análisis de los datos, orientado a obtener las respuestas a las preguntas de investigación planteadas.

Selección de la Muestra

1. Dada la pregunta "¿como afecta el comportamiento de la ciudadanía en Twitter en la agenda pública nacional?", se debe definir que segmento de la agenda sería estudiado, para esta memoria fue acotado a las cuatro principales reformas del gobierno de Bachelet: la Reforma Educacional, Constitucional, Laboral y Tributaria.

2. Para estudiar estos temas particulares en Twitter, se decidió extraer información a través de los hashtags, los cuales son etiquetas utilizadas por los usuarios para referir ciertos temas. Poseen la conveniente propiedad de catalogar temas de conversación (Bruns y Burgess (2011)). Sin embargo, al ser creados por los mismos usuarios pueden existir muchas variantes de hashtags para hablar del mismo tema, además su tiempo de vida es corto dependiendo de cuando el tema en cuestión se encuentra en la palestra de la conversación.

   Para buscar los mejores hashtags se utilizó la herramienta de Google Trends[6], con la cual se pueden buscar los hashtags y temas relevantes (*Trending Topics*), filtrando por zonas geográficas y fechas. Gracias a esta herramienta, se llegó a los diecisiete hashtags mencionados en la tabla 3.1.

[6]https://trends.google.cl/trends/

Estos hashtags han sido categorizados como **oficiales**, los cuales son aquellos que usan las cuentas del gobierno (ministerios y entes oficiales), y **no oficiales**. que son aquellos utilizados por los usuarios tradicionales.

Tipo	Constitucional	Educacional	Laboral	Ttributaria
Oficial	#ReformaConstitucional	#ReformaEducacional	#ReformaLaboral	#ReformaTributaria
No oficial	#ProcesoConstituyente #NuevaConstitucion #AsambleaConstituyente #EncuentrosLocales #UnaConstitucionParaChile #FinAlCae #NosVemosEnLasCalles	#ReformaEducativa #EducacionGratuita #EducacionDeCalidad #NoAlLucro #OfensivaEstudiantil		#ForoTributario

Tabla 3.1: Hashtags por reforma

(Fuente: Elaboración Propia)

Recolección de Datos

1. Utilizando la API de Twitter y Tweepy[7] (Paquete de conexión con dicha API para Python) se desarrolló un programa con el cual, mediante la tecnología de streaming recibe todos los tweets realizados con los hashtags especificados anteriormente. El Streaming permite dejar canales abiertos los cuales reciben el nombre de *listeners*, los cuales escuchan a cualquier dato entrante que concuerde con los filtros dados (Ver ilustración del filtro en figura 3.2). Cada tweet recibido corresponde a una compleja estructura de datos, donde los datos más importantes para esta memoria se pueden ver en la tabla 3.2 e incluyen datos como fecha, texto, autor y número de retweets.

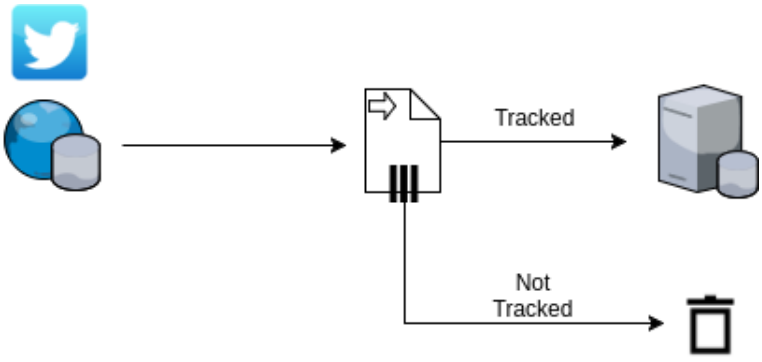


Figura 3.2: Diagrama de captura de Tweets

(Fuente: Elaboración Propia)

[7] http://www.tweepy.org

Dato	Tipo	Descripción
created	Date	Fecha de creación de Tweet
favoriteCount	Integer	Contador de Favoritos del Tweet
id	Long	identificador del usuario creador
isRetweet	Boolean	Indica si el tweet corresponde a un Retweet
latitude	Float	latitud de posición geográfica
longitude	Float	longitud de posición geográfica
retweetCount	Integer	Cantidad de retweets del Tweet
screenName	String	Nombre de usuario en Twitter
statusSource	String	URL del Tweet
text	String	Contenido del tweet

Tabla 3.2: Estructura de un Tweet

(Fuente: Elaboración Propia)

2. En beneficio del análisis de los datos recolectados, se segmentó a los usuarios publicantes en 4 tipos: **Poder Ejecutivo, Poder Legislativo, Prensa Tradicional y ciudadanía**. La lista del poder ejecutivo incluye a ministros, la presidenta y ministerios involucrados en las reformas analizadas; La lista del poder legislativo incluye a diputados y senadores; La lista de la prensa tradicional incluye las cuentas de Periódicos, canales de Televisión y Radio, inscritas en la subsecretaría de Telecomunicaciones de Chile. La lista de ciudadanía son todos los usuarios que twittearon sobre los hashtags seleccionados que no clasifican en las listas anteriores. La tabla 3.3 muestra los usuarios para cada uno de los actores. Se logró identificar las cuentas de Twitter del 95.5 % de todos los senadores y diputados, y 64.5 % del total de medios de prensa registrados en la subsecretaría. Estas faltas de cuentas se debe a que no se encontraron cuentas oficiales de Twitter para estos actores. Sin embargo, algunos poseían paginas web como medio de comunicación.

Actor	Número de usuarios
Poder Ejecutivo	8
Poder Legislativo	151
Prensa Tradicional	347
Ciudadanía	12729

Tabla 3.3: Cuentas obtenidas

(Fuente: Elaboración Propia)

Análisis de datos

En orden a responder las preguntas de investigación planteadas, se realizaron 3 análisis diferentes con los datos recolectados

Análisis Descriptivo:

Se realizaron análisis estadísticos básicos con el objetivo de describir el comportamiento de los datos respecto a frecuencias, comportamiento sobre el tiempo, indices de desigualdad, entre otros indicadores para cada grupo de actores sociales y para cada reforma, separados en los siguientes análisis:

1. Frecuencias de publicación: Es un primer vistazo a los datos, el cual separa el número de tweets, retweets y usuarios por cada reforma y como estos tweets fueron publicados en el tiempo. El objetivo es conocer las reformas dominantes en cuanto a las variables mencionadas y si existieron peaks de publicación en el tiempo.
2. Influencia de actores sociales: Este análisis es el primer paso a responder la pregunta de investigación relacionada a la influencia. Se definieron tres conceptos para medir la presencia de los distintos actores sociales en cada reforma.
   - Emisión: Medida que cuenta la cantidad de tweets emitidos por un usuario o un grupo de usuarios.
   - Alcance: Medida que considera la cantidad de Retweets de una publicación de un usuario o un grupo de usuarios. Esta variable habla del alcance de la información publicada a través de la red gracias a la acción de otros usuarios dentro de la misma.
   - Difusión: Medida que representa la suma de los dos términos anteriores (emisión + alcance). La idea principal de este concepto es conocer una medida de la magnitud en cantidad de información para cada usuario o grupo de usuarios, dado que gracias a la existencia de los retweets, la información publicada por un usuario puede ser replicada. Por ejemplo, una publicación de un usuario B puede ser autoría de un usuario A gracias a un retweet; por lo tanto la información del usuario A se ha publicado dos veces.
3. Clasificación por Hashtags: Para obtener los datos de acuerdo a los tópicos requeridos se utilizó un filtro recolectando los tweets con mención a los hashtags en la tabla 3.1. Al tener hashtags de carácter oficial y no oficial, se quiere saber cual de los hashtags logró traer más datos en la recolección, su comportamiento sobre el tiempo y cuales de los diferentes actores utiliza un tipo determinado de hashtag en volumen.

4. Distribución de retweets por usuarios: Este análisis cuestiona la equidad de la información en Twitter creando distribuciones de retweets, El objetivo es conocer que tanta desigualdad existe en estas distribuciones. Para medir esta desigualdad se utilizó el coeficiente de Gini, el cual es una medida utilizada para analizar distribuciones. Este entrega valores entre 0 y 1, siendo 0 la inexistencia de desigualdad y 1 la mayor desigualdad (Yitzhaki, 1979).

Social Network Analysis:

Se crearon grafos dirigidos para representar las redes para cada una de las reformas.

Para cada red, un usuario corresponde a un nodo de un color según el tipo de actor y un tamaño dependiente de la participación del usuario en base a la cantidad de tweets, mientras que el tamaño del nombre de usuario varía de acuerdo al grado de entrada, es decir, los nombres de usuario mas grandes en la visualización serán aquellos retwitteados por mas usuarios diferentes. En estas redes la relación de retweets mencionada anteriormente entre ambos usuarios corresponde a un arco. Estos arcos, poseen un peso representando interacciones repetitivas (retweets).

Para visualizar cada una de estas redes se utilizó el algoritmo de Fruchterman-Reingold y Force Atlas, algoritmos de distribución de redes que simulan un sistema de fuerzas entre los nodos, donde el objetivo es minimizar la energía total de este sistema, creando una red distribuida homogéneamente (Fruchterman y Reingold, 1991). Estos algoritmos corresponden a la categoría llamada *Force Directed Graph Drawing*.

Para el análisis de cada red se tomó en cuenta métricas como el grado de la red incluyendo sus medias, medianas, maximos, desviaciones estándar y distribuciones, para conocer la cantidad de usuarios con los que interactúa un nodo en la red y la comparación de esta distribución con otras redes. Para conocer que tan aislada se encuentran las componentes en la red se analizaron las métricas de densidad y el coeficiente de agrupamiento (*Clustering Coefficient*). Las componentes conexas por su parte, permiten conocer la cantidad determinada de sub redes no conectadas entre si dentro del grafo y cuales son las componentes más grandes de la red. Una métrica interesante de analizar, es el caso de los conjuntos independientes maximales, se ha utilizado la media entre 50 conjuntos independientes maximales. Este número es debido a que pueden existir muchas formas de crear un conjunto independiente maximal, por lo que se utilizó un número relativamente grande de opciones para estimar el número de vértices. Esta métrica se calculó para tener una medida sobre que porcentaje de los usuarios actúan como conectores principales de la red, es decir, sin estos usuarios la red se rompe.

Esto se explica dado que si la cantidad de vértices en el conjunto maximal es pequeña en relación al total de vértices, significa que la mayoría de usuarios se conocen entre ellos, permitiendo una amplia difusión de la información en la red. Sin embargo, si la cantidad de vértices en el conjunto mencionado es alta, implica que la mayor parte de las personas se relacionan utilizando como puente a una minoría, dando a conocer la

existencia de una élite de usuarios en la red por la que pasa la mayor parte de la información.

Estas redes fueron separadas de acuerdo a su tipo de hashtag (oficial y no oficial), y analizadas de igual forma. Por último se contrasta el comportamiento de las redes de cada reforma, con los diez usuarios mas influyentes de la red. Esta influencia es definida de dos maneras diferentes:

- Influencia temporal: Incluye a los usuarios con mayor media de retweets recibidos por tweet publicado. Esto significa que usuarios con muy pocos tweets, pero muchos retweets en ellos pueden aparecer en esta métrica de influencia. Se le llama temporal dado que son usuarios que en su mayoría tienen su "minuto de fama" con un par de publicaciones que causan mucho impacto. Sin embargo, esta popularidad en la red desaparece rápidamente, por lo que suelen ser usuarios no consolidados dentro de las comunidades en las redes, y por lo mismo, no aparecen en las visualizaciones construidas de las redes.
- Influencia permanente: A diferencia de la influencia temporal, estos usuarios publican constantemente, donde en cada publicación tienen una cantidad constante de retweets superior a la media de retweets de cada usuario, con una base de seguidores fija sobre las cuales se ejerce influencia, y con posibilidad de que esta base de usuarios crezca en el tiempo.

Si un usuario posee una influencia constante, donde en cada publicación posee mas retweets que un usuario con influencia temporal, se está indudablemente ante un usuario influyente en la red, de acuerdo a estas dos métricas.

Análisis de Sentimientos

Por último, para el análisis del contenido de los tweets, se realizó un análisis de sentimientos para determinar la opinión de los usuarios en Twitter respecto a cada una de las reformas y que clase de información es la que posee mayor influencia en la red en base a retweets, para conocer si la información que mas se difunde por la red es de connotación positiva o negativa. Para este caso se utilizará una SVM como método de clasificación de la data, debido a la confiabilidad del método; considerando que se experimentará con datos en español, se requiere una herramienta conocida.

Este clasificador funciona mediante la metodología de entrenamiento y testing, con la ayuda de datos previamente clasificados. Estos datos son separados en dos conjuntos: conjunto de entrenamiento (80 % de los datos) y conjunto de testing (20 % de los datos). Con el primer conjunto se entrena al modelo, pareando distintas entradas con las salidas esperadas, razón por la que usa la mayor parte de los datos etiquetados. El segundo conjunto se utiliza para medir la precisión del modelo posterior al entrenamiento.

Para mejorar este proceso se utiliza un proceso llamado *Cross Validation*, el cual consiste en hacer distintos conjuntos de entrenamiento y testing, a través de distintas combinaciones de los datos, utilizando el que posea las mejores métricas de precisión.

Los datos fueron clasificados para separar tweets positivos y negativos. Cabe destacar que no se incluyó la categoría neutral, dada la poca data de entrenamiento etiquetada en esta categoría, lo cual crea clases desbalanceadas, afectando la precisión de la predicción.

Para el entrenamiento y testing de la SVM, se trabajó con datos pertenecientes a la Sociedad Española para el Procesamiento del Lenguaje Natural, organización que realiza cada año el TASS[8], los cuales son *workshops*, dedicados a realizar análisis de Sentimientos en idioma español. Dicha organización facilitó sus *datasets* etiquetados en tópicos políticos para fines académicos. La data de entrenamiento utilizada posee un 53.994 % de data asociada a un sentimiento positivo y un 46.006 % de data catalogada como sentimientos negativos.

Para realizar el procedimiento, se utilizó una máquina virtual CentOS, creada dentro de un cluster, con 16 GB de Ram, tomó 2 días en encontrar los mejores parámetros de optimización del modelo, obteniendo una precisión de entrenamiento y testing de 84.1 % y 80.4 % respectivamente, respecto a los datos del TASS. La tabla 3.4 muestra los parámetros, sus opciones y el valor elegido luego de esta búsqueda que optimiza el modelo.

Parámetro	**Opciones**	**Elección**
Rango de n-gramas	[Unigramas, Bigramas]	Bigramas
C	[0.2,0.5,0.7]	0.2
Función de perdida	[hinge, squared hinge]	Hinge

Tabla 3.4: Búsqueda de parametros SVM

(Fuente: Elaboración Propia)

[8]http://www.sepln.org/workshops/tass/2016/tass2016.php

- **Procedimiento de clasificación**

  1. *Cleaning*: La limpieza de cada tweet corresponde a reconocer los números y símbolos, para no incluir signos de pregunta o exclamaciones como palabras aisladas en los procesos siguientes.
  2. *Stemming*: El proceso llamado *Stem*, en inglés, transforma cada palabra encontrada a su forma raíz, por ejemplo, las palabras esperaba, espera, esperando, corresponden a la palabra raíz: esperar. El objetivo de realizar esto es disminuir el número de variables en el modelo, mejorando su eficiencia y evitando que ciertas variaciones de una palabra tengan un peso mayor por sobre otras.
  3. *Tokenizing*: Cada palabra encontrada posterior al *Stemming*, es asignada a un token que la identifica.
  4. Binarización: La data es asignada en valores binarios (0 o 1), si es negativa o positiva.
  5. Búsqueda de parámetros: Para tener una buena aproximación en el modelo, antes de ejecutarlo se realiza una búsqueda de parámetros para encontrar la mejor configuración dados los datos. Esta búsqueda se realiza ejecutando el modelo con las distintas combinaciones entre los candidatos a parámetros y probando los resultados con los datos de testing. Cabe resaltar que mientras más parámetros y data otorgamos a la búsqueda, más tarda en encontrar los mejores parametros.
  6. Entrenamiento: Con los parámetros encontrados, se procede a la creación del modelo y la prueba de la clasificación utilizando la data de testing.
  7. Testing: Con el modelo construido se clasificó la data obtenida de los tweets para cada reforma. Cabe destacar que la data de entrenamiento fueron tweets con contexto político, al igual que los datos recolectados. Esto agrega confianza al modelo dado que los datos utilizados en el entrenamiento no solo están en el mismo lenguaje, sino que poseen la misma estructura, y el mismo contexto lingüístico.

- **Validación del modelo**

  Para validar el modelo se obtuvo un subconjunto aleatorio del 10 % de los datos. Este subconjunto fue sometido a una clasificación de sentimiento de manera manual por dos anotadores diferentes quienes no pudieron ver las respuestas del otro. Posteriormente esta muestra se clasificó en la SVM usando los resultados de la clasificación manual como “valor real” para obtener las métricas de accuracy, recall y la matriz de confusión para los datos clasificados y sus etiquetas.

Para determinar la confiabilidad de la clasificación manual, se calculó el **Coeficiente Kappa de Cohen**, también conocido como *Intercoder reliability Kappa* (Wood, 2007). Este estadístico mide la concordancia entre dos anotadores en la clasificación de un mismo conjunto de datos. La tabla 3.5 muestra los valores del Kappa de Cohen para catalogar el nivel de concordancia entre dos anotadores según (Landis y Koch, 1977).

Kappa	**Nivel de concordancia**
$\kappa < 0$	No acuerdo
$0 \leq \kappa \leq 0{,}20$	Acuerdo Leve
$0{,}21 \leq \kappa \leq 0{,}40$	Acuerdo Razonable
$0{,}41 \leq \kappa \leq 0{,}60$	Acuerdo Moderado
$0{,}61 \leq \kappa \leq 0{,}80$	Acuerdo Substancial
$0{,}81 \leq \kappa < 1{,}00$	Acuerdo Casi Perfecto

Tabla 3.5: Clasificación del Kappa de Cohen
(Fuente: Elaboración Propia)

Este coeficiente se calcula de la siguiente forma:

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \tag{3.6}$$

Donde:

- p_0: proporción de acuerdo entre anotadores. Valor entre [0, 1]
- p_e: probabilidad hipotética de estar de acuerdo por azar.

$$p_e = \frac{1}{N^2} \sum_k n_{k1} n_{k2} \tag{3.7}$$

  - N: Número de items a clasificar
  - n_{ki}: Número de veces que el anotador i prefirió la categoría k

Tecnología utilizada

Para recolectar la data se utilizó la API de Twitter mediante la biblioteca tweepy[9] para Python, la cual provee el sistema de Streaming para la captación de tweets.

Para realizar el análisis descriptivo y de redes se utilizaron principalmente las bibliotecas Pandas, Numpy, Matplotlib y Networkx, todas bibliotecas de código abierto. Cada una de ellas es descrita a continuación:

- Pandas[10] permite convertir la data en dataframes para facilitar su análisis mediante filtros y obtención de datos "like-SQL" . Es decir, mediante un lenguaje de consultas acceder a diferentes datos según ciertos parámetros.
- Numpy[11] es una poderosa biblioteca para realizar cálculos numéricos y trabajar con álgebra lineal, tanto con escalares, vectores y matrices, permitiendo realizar variados procedimientos de forma optimizada.
- Matplotlib[12] es una herramienta de visualización muy utilizada para python que permite realizar gráficas y diferentes tipos de visualizaciones y animaciones.
- Networkx[13] es una biblioteca de creación y análisis de grafos la cual posee estructuras de datos propias para representarlos, permitiendo crear redes propias, manipularlas, aplicar diversos algoritmos de la teoría de grafos y obtener métricas de estas redes.
- Gephi[14], por último, es un software de construcción y análisis de redes open source, utilizado para crear y visualizar las redes.

Para este análisis se utilizaron principalmente las bibliotecas NLTK y Scikit-learn, ambas de código abierto para Python.

- NLTK[15] (Natural Language Toolkit) es un paquete de procesamiento de lenguaje humano el cual permite clasificar, tokenizar, analizar estructura lingüistica y manipular en diferentes sentidos data de lenguaje.
- Scikit-Learn[16] es una biblioteca dedicada a algoritmos relacionados con *Machine Learning*, facilitando herramientas para *data mining* y análisis de datos, gracias a algoritmos de clasificación, preprocesamiento, reducción de dimensionalidad, redes neuronales, entre otros.

[9] http://www.tweepy.org/
[10] http://pandas.pydata.org/
[11] http://www.numpy.org/
[12] https://matplotlib.org/
[13] https://networkx.github.io/
[14] https://gephi.org/
[15] http://www.nltk.org/
[16] http://scikit-learn.org/

Resultados

Para presentar los resultados se utilizará la metodología propuesta de análisis de datos. Para este capítulo, se separan los resultados de acuerdo a los distintos tipos de análisis utilizados: Análisis descriptivo, Análisis de redes y Análisis de sentimientos.

Análisis Descriptivo

El primer análisis es un análisis descriptivo de los datos basado en frecuencias con el objetivo de encontrar patrones que puedan describir las tendencias y comportamientos de los usuarios en Twitter.

Frecuencias de publicación

Se conoce a primera vista que en ciertos periodos de tiempo, temas importantes salen a la palestra y se convierten por unos momentos en temas sumamente debatidos en Twitter. Por lo tanto, lo primero que se debe conocer es la magnitud de los datos y como se distribuyen estos en el tiempo de recolección, el cual fue desde el 1 de Julio al 31 de Agosto del 2016. Este periodo de recolección basado en los hashtags determinados para la selección de la muestra arrojó aproximadamente treinta y cuatro mil tweets, los cuales fueron retweeteados más de ochocientas mil veces, solo por un total de aproximadamente trece mil usuarios diferentes.

La tabla 4.1 muestra que las reformas Constitucional y Educacional poseen la mayor parte de los tweets con un valor del 47 % y 29 % respectivamente respecto al total en la tabla, en comparación con las reformas Laboral y Tributarial, las cuales poseen un 14 % y 9 % respectivamente, superando estas dos primeras reformas en aproximadamente un 300 % a las últimas dos. Esto puede deberse a diversos motivos, como cobertura en la prensa, niveles de tecnicismo del tópico en cuestión, percepción de los ciudadanos sobre como los afecta cada tópico, etc.

Uno de los números mas interesantes de analizar es la cantidad total de usuarios que publicaron sobre estos temas, el cual es un número mas pequeño, en comparación al total de ciudadanos en Chile (13235 contra aproximadamente 18 millones, según proyecciones del INE[17]). Este número también se compara con los datos publicados por Adimark, donde el 56 % de las personas encuestadas declara conocer Twitter, sin embargo solo un 19 % de estos declara usarlo.

		Twitter	
Reform	**Tweets**	**Retweets**	**Users**
Constitutional	16,281	221,414	6,019
Educational	9,971	379,618	4,337
Labor	4,901	180,043	3,019
Tax	3,221	37,664	1,248
Total	34,374	818,739	13,235

Tabla 4.1: Resumen descriptivo de los datos por reforma

(Fuente: Elaboración Propia)

Estos valores permiten deducir en primera instancia que solo una pequeña parte es la creadora del contenido que se difunde por la red, junto a un volumen desconocido que actúan como lectores pasivos (es decir, no realizan retweets ni publican por ellos mismos). Con esta cantidad de tweets, el siguiente paso es conocer la distribución en los dos meses de observación de estos.

Como se puede ver en la figura 4.1, en 62 días correspondientes al periodo Julio-Agosto del 2016 los principales peaks en aumentos se encuentran en los días centrales, es decir, entre el 20 de Julio y el 15 de Agosto aproximadamente, fechas en las cuales se discutieron fuertemente noticias relacionadas a las reformas educacional y constitucional. Ejemplos de esto pueden verse en la tabla 4.2, donde se mencionan eventos publicados por los medios de prensa tradicionales, en las fechas con alzas vistas en la figura 4.1.

Para las reformas laboral y tributaria existe presencia de peaks sin noticias offline, por lo que se requiere un análisis mas acabado de los datos para explicar este comportamiento.

[17] http://www.ine.cl/estadisticas/demograficas-y-vitales

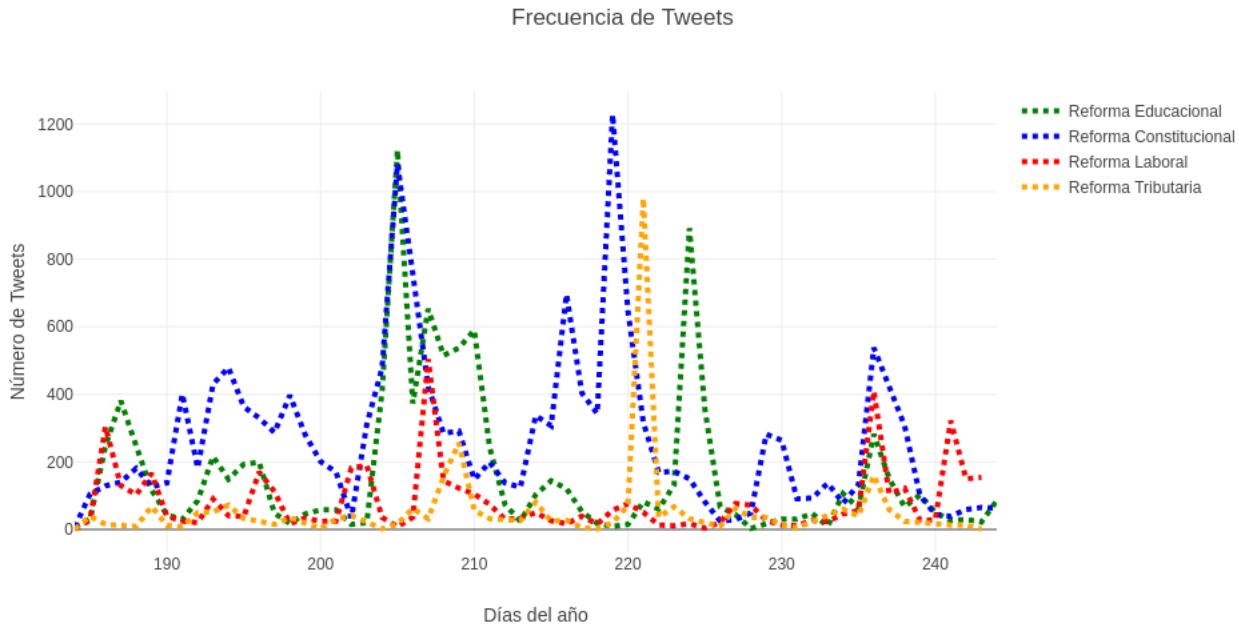


Figura 4.1: Tweets por día, con eventos offline

Elaboración propia

Reforma Constitucional	**Noticia**
21-07-2016	Diputados proponen reforma constitucional que termina con las AFP
07-08-2016	Fin del proceso de cabildos regionales
Reforma Educacional	**Noticia**
23-07-2016	Convocatoria a marcha Nacional Estudiantil
27-07-2016	Declaraciones del G9 ante proyecto de reforma Educacional
27-07-2016	Comienza el caso Roxana Pey (Rectora despedida de la universidad de Aysen)

Tabla 4.2: Muestra de eventos offline en fechas de peaks online

(Fuente: Elaboración Propia)

Influencia de actores sociales en Twitter

El objetivo de este análisis es conocer de entre los datos recolectados cuál de todos los actores entre el Poder Legislativo, Poder Ejecutivo, Medios de prensa y Ciudadanos posee mayor dominio sobre la información generada en Twitter, mediante los conceptos definidos en la metodología como emisión, alcance y difusión.

- Emisión: Como se puede ver en la figura 4.2, expresada en escala logarítmica, se puede ver que casi la totalidad de Tweets provienen de la ciudadanía, con más del 99 % para cada una de las reformas, mientras que los otros actores emiten menos del 0.5 % de los tweets captados o simplemente no crean contenido. Para el caso de la reforma laboral y tributaria, se detecta la presencia de los actores políticos y de prensa, con un valor muy similar a las dos principales reformas, lo que quiere decir que existen grupos pertenecientes al poder legislativo, ejecutivo y prensa chilena que efectivamente hablan de las reformas bajo estudio.

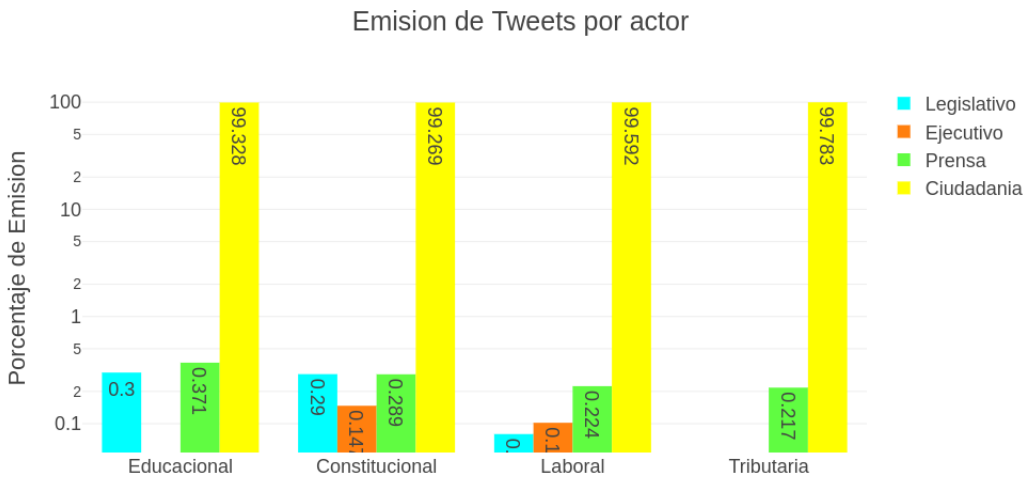


Figura 4.2: Emisión de Tweets por actor

Fuente: Elaboración propia

- Alcance: En este caso, la figura 4.3 muestra que el poder legislativo aumenta su participación en las reformas educacional, constitucional y laboral, al ser responsable del 10 % de retweets para la reforma educacional, un 4 % para la reforma constitucional y un 1 % para la reforma laboral, mientras que la reforma tributaria no posee participación de este actor.

  Para el caso del poder ejecutivo, el cual solo tiene presencia en las reformas constitucional y laboral, no aumenta demasiado su participación respecto a la figura 4.2.

  La prensa tradicional por su parte, si bien tiene representación en cada una de las reformas, posee una participación menor en cuanto a responsabilidad de retweets producidos.

  La ciudadanía se mantiene como el actor social con mas participación y producción de contenido dentro de todas las reformas.

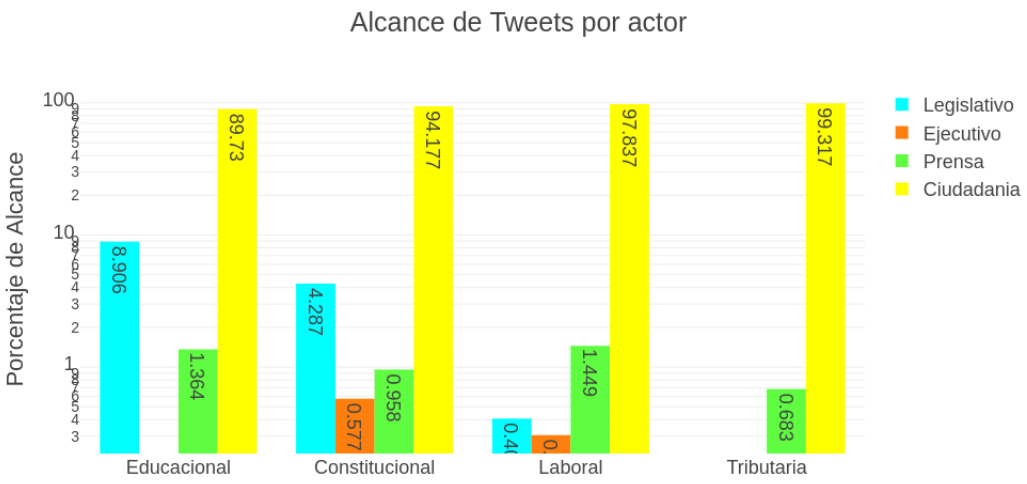


Figura 4.3: Alcance de Tweets por actor

Fuente: Elaboración propia

- Difusión: La figura 4.4 muestra incrementos en todos los poderes y los medios de prensa. Sin embargo, se observa una baja en la ciudadanía. Esto a simple vista revela que el potencial de difusión de información de la ciudadanía es menor que los otros tres actores, a pesar de mantener el dominio de publicaciones en términos numéricos.

  Se puede concluir que ante la diminuta cantidad de contenido publicado por el poder legislativo, el potencial de diseminación por la red es el más alto para las reformas educacional y constitucional con un factor de incremento de 39 y 14 respectivamente (ver tabla 4.3).

Distinto es el caso para el poder ejecutivo y los medios de prensa, para los cuales su participación creció en un factor de cinco aproximadamente, valor mucho menor que el caso anterior. Por otro lado, la prensa es el único actor social (además de la ciudadanía) que está presente en cada una de las reformas lo que lo hace un actor mas transversal a los temas bajo estudio en comparación a ambos poderes del estado.

En conclusión, y relacionado a las preguntas de investigación, este análisis arroja luces de que si bien la ciudadanía posee la mayor cantidad de publicaciones y presencia en la red, el poder legislativo posee una mayor capacidad de difusión, existiendo un dominio por parte de estos dos actores en la red, mientras que la prensa mantiene una cobertura general a todos los tópicos sin poseer una difusión importante, por lo que se necesitan más análisis para responder si esta posee algún control de la información.

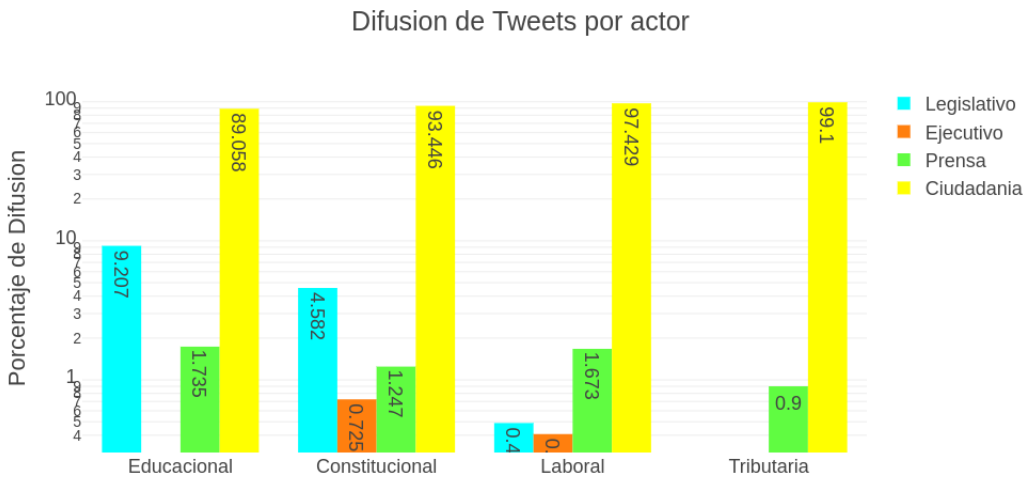


Figura 4.4: Difusión de Tweets por actor

Fuente: Elaboración propia

	Emisión				Difusión			
	Legislativo	**Ejecutivo**	**Prensa**	**Ciudadanía**	**Legislativo**	**Ejecutivo**	**Prensa**	**Ciudadanía**
Educacional	0.301		0.371	99.328	8.906		1.364	89.730
Constitucional	0.294	0.147	0.289	99.269	4.287	0.577	0.958	94.177
Laboral	0.082	0.102	0.224	99.591	0.408	0.306	1.449	97.837
Tributaria			0.217	99.782			0.683	99.317

Tabla 4.3: Porcentajes de emisión y difusión de contenido

Fuente: Elaboración propia

Ruido en los Datos Recolectados

En la estructura de un tweet, existen campos de latitud y longitud que permiten filtrar la geolocalización de un usuario al momento de publicar un tweet. Sin embargo, al momento de analizar los datos, se pudo observar que estos campos aparecieron vacíos en el 99.9 % de los tweets en cada una de las reformas. Esto permite que tweets creados usando un hashtag idéntico, puedan pertenecer a localidades diferentes, con contextos sociales parecidos (por ejemplo, una reforma tributaria diferente). Ante esta potencial fuente de error, se utilizó una estimación basada en la ubicación proporcionada por los usuarios en su información personal, evaluando si el campo de ubicación está en el país, es vacío o pertenece al extranjero. Si bien esta variable no es tan certera como la geolocalización, debido a que puede existir información falsa, puede ayudar a entender mejor los datos.

La tabla 4.4 muestra que para las reformas laboral y tributaria, el 40 % y 60 % respectivamente declaran ser usuarios extranjeros, y a su vez, poseen pocos campos vacíos, dando poco espacio a que estos datos faltantes sean capaces de darle confiabilidad al conjunto de tweets.

Las reformas educacional y constitucional por su parte, poseen menor proporción de usuarios que se declaran extranjeros, con 35 % y 23 % respectivamente. Esto, sumado a la gran cantidad de datos vacíos, hace necesario realizar más análisis sobre estas dos reformas. Por lo tanto, debido a la baja cantidad de tweets y al alto nivel de ruido, se descartarán las reformas laboral y tributaria de los análisis posteriores, debido a que el nivel de ruido aporta negativamente a sus resultados.

	Twitter		
Reforma	**Chile**	**Extranjero**	**Vacío**
Constitucional	41.42 %	23.80 %	34.78 %
Educacional	24.95 %	35.01 %	40.04 %
Laboral	43.59 %	40.95 %	15.46 %
Tributaria	26.90 %	61.62 %	11.48 %

Tabla 4.4: Análisis de ruido en los datos

(Fuente: Elaboración Propia)

Clasificación por Hashtags

La figura 4.5, muestra las proporciones de uso de los distintos tipos de hashtag por cada actor, donde se puede ver que todos los actores utilizan ambos tipos de hashtags, superando levemente el uso de hashtags oficiales (ver tabla 4.5). Para el caso de la ciudadanía, el uso de hashtags oficiales es similar en ambas reformas. Sin embargo para el resto de los actores, se puede ver que el uso de hashtags no oficiales es menor con la excepción del poder legislativo en la reforma constitucional, siendo este uso nulo para el caso del poder ejecutivo y con mayor diferencia para los medios de prensa, posiblemente debido a la objetividad que deben mantener estos.

A partir de la tabla 4.5, una observación importante es que a pesar que la mayor parte de las publicaciones pertenecen a la ciudadanía, el uso de hashtags oficiales v/s no oficiales es similar en ambas reformas, siendo levemente mayor el uso de hashtags oficiales, lo cual significa que los usuarios en Twitter utilizan tanto los hashtags creados en las campañas del gobierno como los creados por otros usuarios para manifestarse respecto a estos temas. Esto muestra el dinamismo de los hashtags en Twitter, donde los usuarios pueden usar el medio oficial, para llegar a los creadores del hashtag (en este caso el gobierno), como también usar medios no oficiales para demostrar opiniones o converger ciertos temas.

Reforma	**Hashtag Oficial**	**Hashtag No Oficial**
Constitucional	7657	6689
Educacional	2950	2555

Tabla 4.5: Tabla resumen: Número de publicaciones según tipo de hashtag

(Fuente: Elaboración Propia)

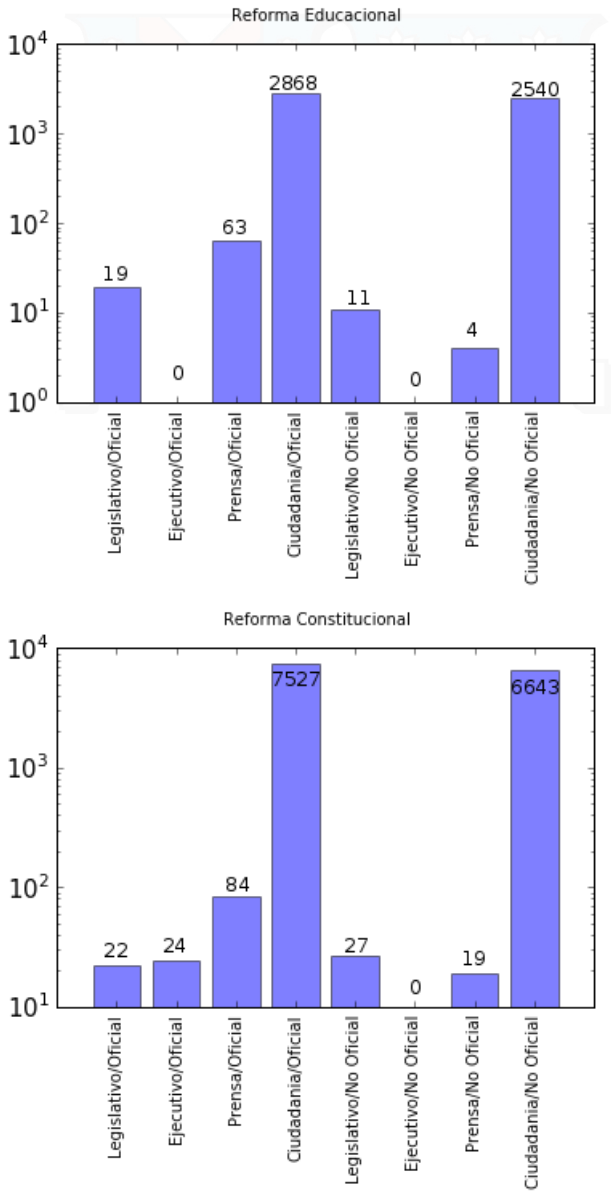


Figura 4.5: Clasificación de Hashtags por reforma

Fuente: Elaboración propia

Ante el resultado obtenido donde el uso de hashtags es muy similar, se realizó una línea temporal para descubrir como es el comportamiento del uso de hashtags oficiales y no oficiales a lo largo del periodo de recolección. La figura 4.6 muestra que aproximadamente el primer 60 % del tiempo de recolección el uso de hashtag no oficiales en la reforma educacional era en general menor, utilizándose más el hashtag oficial para publicar acerca de esta reforma. Sin embargo, en los últimos 20 días el uso de hashtags no oficiales desplazó a los oficiales. Esto coincide con los eventos ocurridos en esos días en la tabla 4.2, específicamente los eventos relacionados al caso Roxana Pey y las declaraciones del G9. Esto índica que en el momento en que la movilización estudiantil cobró fuerza mediática, los usuarios optaron por usar sus propios hashtags para debatir y opinar acerca de la reforma educacional, desplazando al hashtag oficial.

Distinto es el caso para la reforma constitucional, donde la figura 4.7 muestra que el uso de los hashtags oficiales y no oficiales van casi a la par, con peaks recurrentes, a diferencia de la reforma educacional, la cual mantenía peaks de publicación mas bajos y solo en el hashtag oficial.

Estos dos casos permiten sospechar que ante movilizaciones sociales o tópicos conflictivos a nivel social, los usuarios tienden a usar hashtags impulsados por ellos como ciudadanía, mientras que para tópicos donde exista menos conflicto, tienden a usar los hashtags oficiales del tópico.

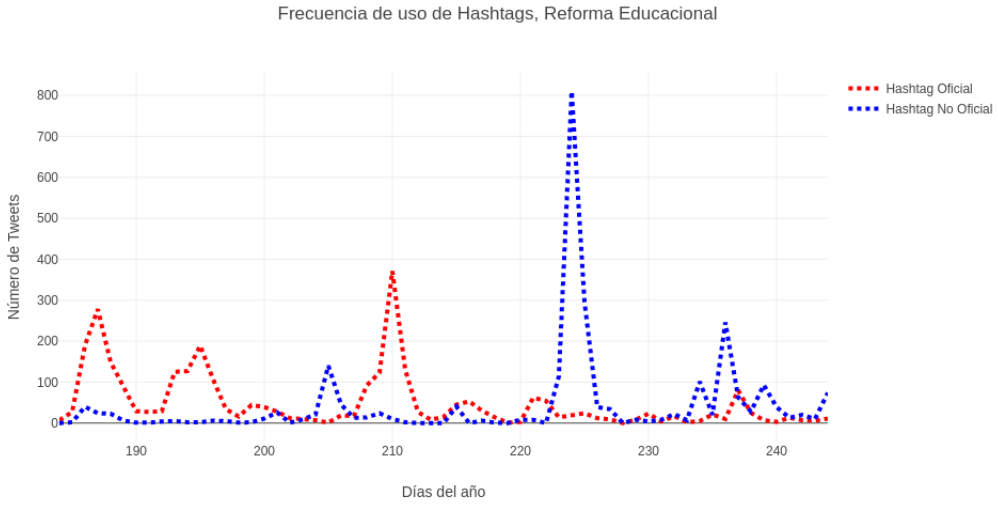


Figura 4.6: Uso de hashtags oficiales en el tiempo: Reforma Educacional

Fuente: Elaboración propia

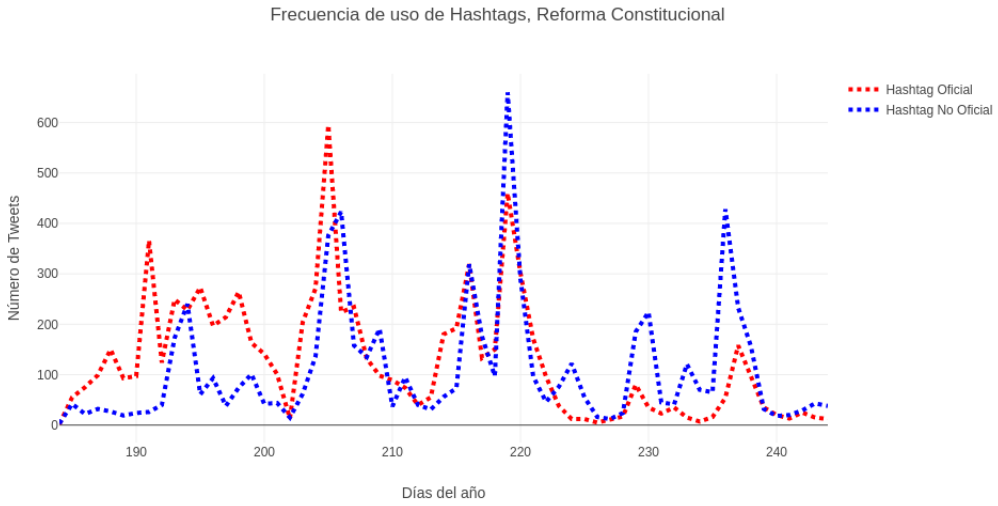


Figura 4.7: Uso de hashtags oficiales en el tiempo: Reforma Constitucional

Fuente: Elaboración propia

Distribución de retweets por usuarios

Una de las características de Twitter es la capacidad de cualquier usuario de dar su opinión sobre algún tema en particular, dando la idea de que existe una equidad en la distribución de la información, donde es equiprobable que un usuario lea la opinión de cualquier otro usuario. Para comprobar esto en el contexto elegido, se ha graficado la distribución de retweets para cada usuario, para conocer si existe esta equidad en la información de cada usuario.

La figura 4.8 toma a todos los usuarios por cada reforma y los ordena en orden descendente respecto al número de retweets recibidos. Al costado de cada curva puede verse su coeficiente de Gini, el cual mide la desigualdad en la distribución. Ambas reformas, poseen una alta desigualdad con un coeficiente de Gini mayor a 0.6, (ver figura 4.8). Esto considerando que al menos en el contexto de ingresos, según las naciones unidas un indice superior a 0.4 es señal de alarma (Un-Habitat, 2008), significa que un valor como los vistos aquí son preocupantes cuando se habla de desigualdad.

La figura 4.9 separa la distribución anterior en tres conjuntos: el 1 % más retwitteados, el siguiente 9 % y el 90 % restante, y muestra la mediana de retweets de cada grupo. Gracias a esta separación se puede ver que los usuarios pertenecientes al 1 % mas retwiteado poseen aproximadamente el doble de retweets que un usuarios perteneciente al 9 % con mas retweets y más de diez veces más que el 90 % de usuarios restantes para la reforma educacional, mientras que para la reforma constitucional el primer grupo supera por el triple al segundo grupo y por casi treinta veces más al tercer grupo, mostrando que existen minorías cuya información viaja con mayor facilidad por la red.

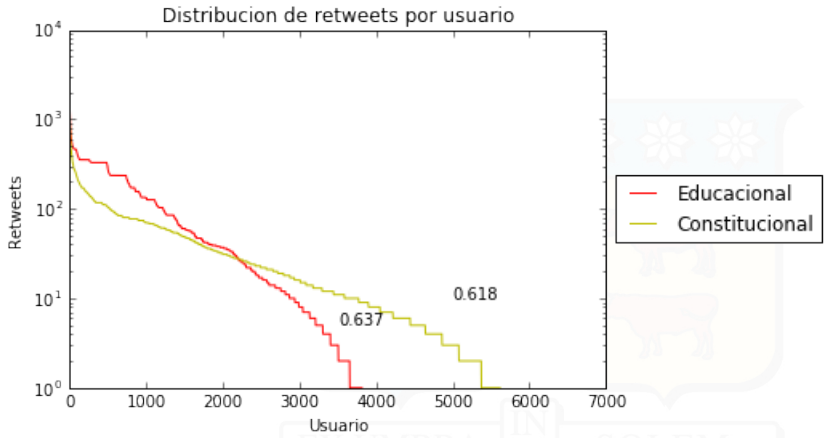


Figura 4.8: Distribución de Retweets por usuario

Fuente: Elaboración propia

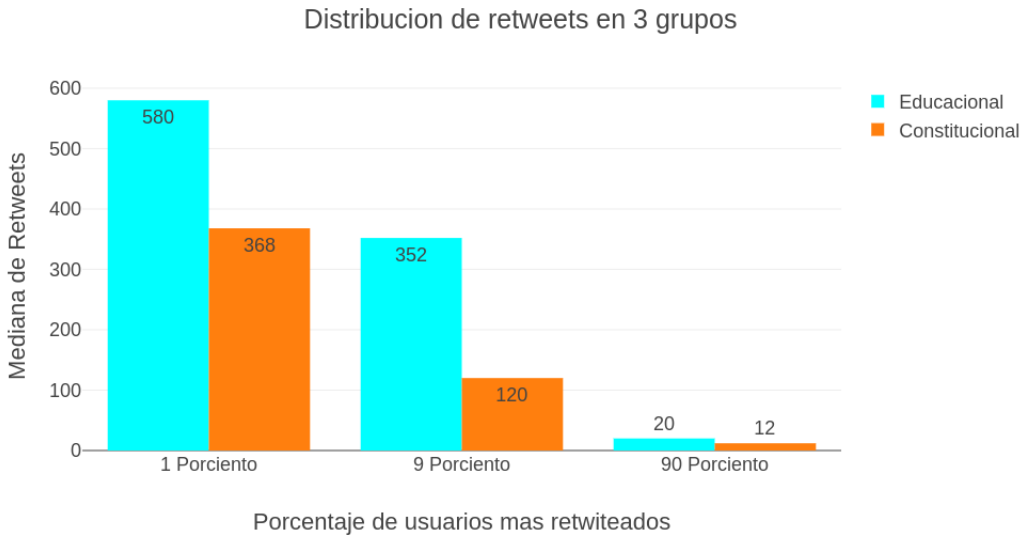


Figura 4.9: Usuarios influyentes e impacto en la red

Fuente: Elaboración propia

Por lo tanto, en el contexto de las reformas estudiadas y respondiendo otra de las preguntas de investigación (**¿Están las redes sociales realmente dando libertad y visibilidad de las opiniones de cada usuario por igual?**), la distribución de la información no es equitativa, concentrándose la mayor parte de la información generada en unos pocos usuarios fuentes, lo que indica la existencia de **usuarios influyentes**, los cuales son usuarios cuyas publicaciones generan un mayor número de retweets en comparación al resto de los usuarios, en este caso el 10 % de los usuarios que twittearon utilizando alguno de los hashtags de las reformas poseen mayor influencia en la red mientras que las publicaciones del otro 90 % no resuenan en de la misma forma (ver figura 4.9)

Conclusiones generales

Mediante esta primera aproximación a conocer los datos recabados, se logra obtener los siguientes puntos notables:

1. Considerando las reformas estudiadas, la reforma constitucional y educacional se llevan la mayor parte de toda la data, acaparando el 76,3 % del total de datos. Esto sumado al hecho de que la mayoría de las publicaciones pertenecen a la ciudadanía, permite especular que estos temas eran prioritarios para los ciudadanos en el lapso de tiempo estudiado.

2. La mayor parte de la información que se difunde por la red pertenece a la ciudadanía. Sin embargo, esto no quiere decir que en los otros tipos de actores no existan usuarios influyentes, dado que la proporción de usuarios para cada uno de los actores es muy diferente. En este caso se tiene una ciudadanía que abarca la mayor cantidad de publicaciones, pero sin la capacidad de difundir su información, a diferencia de los otros actores cuyas publicaciones son capaces de difundirse a través de otros usuarios.

3. Mas allá de conocer al tipo de actor mas presente entre los datos, conocer la cantidad de retweets promedio de los usuarios se vuelve importante para tener una estimación de la conducta de retweeteo de cada uno de estos. En esta linea, sin tomar en cuenta a que clase de actores pertenecen los usuarios, se puede ver que la distribución de retweets por usuarios es bastante desigual, llegando al punto donde el 10 % controla la mayor parte del contenido que viaja por la red, revelando la presencia de usuarios con mayor influencia y mayor control sobre la información que otros.

Análisis de redes

Para este análisis, los datos fueron sometidos a un análisis de redes, creando grafos dirigidos de usuarios (ver figuras 4.10 y 4.11), con el cual conocer el comportamiento o las propiedades de estos al interactuar entre ellos. Estas redes poseen sus vértices de un tamaño dependiente del número de Tweets y un tamaño de Nombre de usuario dependiente del grado de entrada.

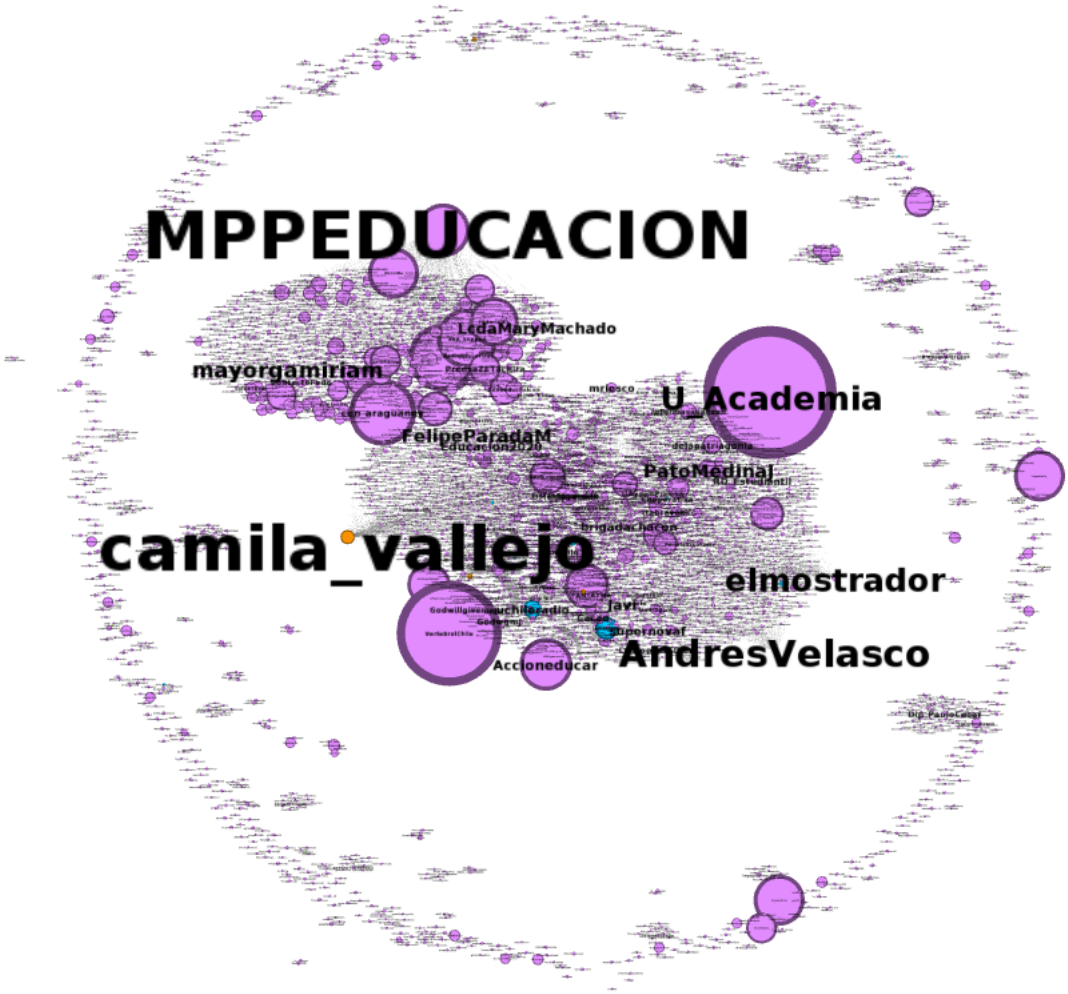


Figura 4.10: Red de la reforma educacional

Fuente: Elaboración propia

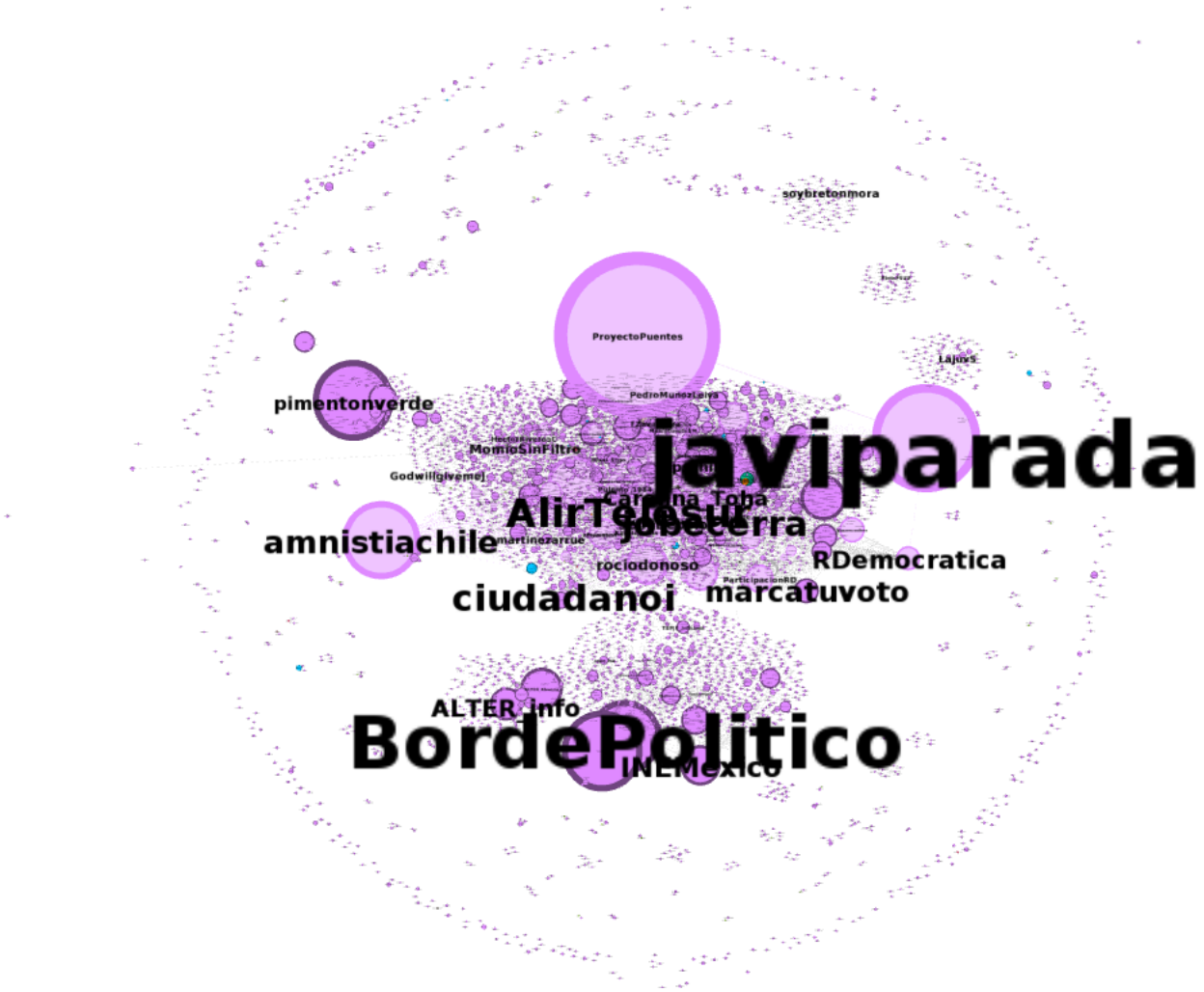


Figura 4.11: Red de la reforma constitucional

Fuente: Elaboración propia

Redes Reforma Educacional y Constitucional

El siguiente análisis por cada métrica se realiza considerando las redes construidas utilizando los datos recolectados, la tabla resumen 4.6 muestra las métricas más importantes para ambas redes.

	R. Educacional	**R. Constitucional**
Vertices	4348	6045
Arcos	4591	8158
Peso de Arcos		
Media	1.482	1.478
Min	1	1
Max	54	36
Desviación Estándar	1.785	0.132
Mediana	1	1
Métricas Generales		
Grado	1.784	2.135
Coef. Agrupamiento	0.0058	0.1054
Densidad	0.0006	0.0005
Componentes Conexas	628	549
Conj. Independiente Maximal	93.93 %	91.48 %

Tabla 4.6: Reforma Educacional y Constitucional: Tabla resumen

(Fuente: Elaboración Propia)

- Grado: El grado de la red describe con cuantos usuarios en promedio interactúa cada usuario. En este caso la tabla 4.6 muestra que cada usuario interactúa en promedio con 2 usuarios mas, lo que significa que cada publicación de un usuario, es potencialmente capaz de difundirse duplicándose en cada nivel. Sin embargo, si se analizan en profundidad los grados para cada actor, se puede ver que existe una diferencia entre el comportamiento de los grados para cada uno. La tabla 4.7 muestra que en el caso de los actores pertenecientes al poder ejecutivo y legislativo mantienen una media y una mediana mucho mayor que los ciudadanos, dado que los usuarios pertenecientes a estos grupos poseen mas visibilidad mediática, lo que complementa los análisis de alcance y difusión de contenido, donde la ciudadanía aparece por sobre estos poderes, gracias a la gran diferencia en número de usuarios. Este análisis de grados permite concluir que los usuarios del poder legislativo y ejecutivo, a pesar de ser pocos en número, poseen una influencia alta en la red, especialmente el legislativo en la reforma educacional.

Reforma Educacional				
Actor	**Media**	**Mediana**	**Desviación Estándar**	**Maximo**
Ejecutivo	-	-	-	-
Legislativo	61.42	9.00	135.04	392.00
Prensa	16.89	4.00	19.75	59.00
Ciudadanía	1.99	1.00	9.86	409.00
Reforma Constitucional				
Actor	**Media**	**Mediana**	**Desviación Estándar**	**Maximo**
Ejecutivo	19.25	16.50	15.20	43.00
Legislativo	10.88	3.50	15.19	54.00
Prensa	2.73	1.00	5.50	27.00
Ciudadanía	2.66	1.00	9.43	359.00

Tabla 4.7: información sobre grados de la red para cada actor y cada reforma

(Fuente: Elaboración Propia)

Para conocer si no existen diferencias entre las distribuciones de los grados de cada reforma, se realizó un T-test, con las siguientes hipótesis:

- H_0: No existen diferencias en el comportamiento de los grados entre la reforma educacional y constitucional
- H_1: Si existen diferencias en el comportamiento de los grados entre la reforma educacional y constitucional

En este análisis, bajo el valor *p-value*< 0,05, existe evidencia suficiente para rechazar H_0 (ver tabla 4.8), por lo que se acepta que ambas distribuciones son diferentes entre las reformas, a pesar de poseer valores de media similares entre ellas, lo cual complementa el análisis por actor social realizado anteriormente para los grados.

	Reforma Educacional
Reforma Constitucional	-2.755 / p = 0,005
p < 0,05	

Tabla 4.8: T-test en distribución de grados

(Fuente: Elaboración Propia)

A partir de estos datos, se analizaron los 10 usuarios con el mayor grado, esto quiere decir que estos usuarios son los que fueron retwitteados por más usuarios distintos en la red. Estos usuarios son mostrados en la tabla 4.9, donde se pueden ver figuras conocidas en el mundo de la política, como Camila Vallejo, Andrés Velasco, cuentas de partidos políticos como Revolución Democrática y solo un medio de prensa, El Mostrador, el cual es un medio no catalogado en medios televisivos, de radio, o de periódicos, ya que pertenece a la categoría de periódicos digitales. Estos usuarios son los que potencialmente pueden interactuar con mas usuarios en la red, siendo éste grupo los principales usuarios sobre los que influyen éstos perfiles en twitter.

Sin embargo, esto también dice que tanto el poder ejecutivo, legislativo y la prensa poseen una influencia mayor al promedio de los ciudadanos dentro de la red, incluso cuando el usuario con el grado mayor pertenece a la ciudadanía para la reforma constitucional, este corresponde a un personaje mediático que no encaja en los otros actores sociales definidos para esta investigación, por lo tanto no es un ejemplar representativo de la ciudadanía.

Se puede observar en la tabla 4.9 que aparecen usuarios pertenecientes al extranjero debido al ruido en los datos, sin embargo al observar las figuras 4.10 y 4.11, se puede ver como el algoritmo separa las comunidades, aislando de la gráfica el conjunto de usuarios que pertenece a un contexto social diferente para ambas reformas.

R. Educacional			**R. Constitucional**		
Usuario	**Rol**	**Grado**	**Usuario**	**Rol**	**Grado**
MPPEDUCACION	Ministerio (E)	409	javiparada	Activista	359
camila_vallejo	Diputada	392	BordePolitico	Colectivo (E)	224
U_Academia	Universidad	253	jobecerra	Periodista	155
AndresVelasco	Ministro	221	AlirTelesur	-	152
elmostrador	Periodico Digital	186	ciudadanoi	Ong	139
mayorgamiriam	Profesora (E)	123	marcatuvoto	Colectivo	138
PatoMedinaJ	Dirigente Estudiantil	106	amnistiachile	Colectivo	129
FelipeParadaM	Militante	98	INEMexico	Institución (E)	123
LcdaMaryMachado	Militante (E)	83	RDemocratica	Partido Político	119
_____Javi	-	75	ALTER_info	Partido Político (E)	113

(E) : Extranjero

Tabla 4.9: Usuarios con mayor grado en cada reforma

(Fuente: Elaboración Propia)

- Peso de los arcos: Los datos indican que dos usuarios cualquiera en la red, en promedio y mediana interactúan 1 vez, para ambas reformas. Sin embargo, existen en la reforma educacional casos aislados de usuarios que interactúan constantemente (ver máximos en tabla 4.6). En la reforma constitucional esto también se observa, pero con un máximo menor. Esto continúa revelando una desigualdad en la interacción entre cada usuario en la red.

- Coeficiente de Agrupamiento: El coeficiente de agrupamiento de la red de la reforma educacional es 0.0058, mientras que en la constitucional es de 0.1054, siendo ambos valores muy pequeños, lo que quiere decir que las comunidades en la red están muy separadas entre si. Esto implica que existen muchos usuarios los cuales no interactúan vía retweets, disminuyendo la probabilidad de que la información creada en una comunidad pueda llegar a otras comunidades a través de mas usuarios. Según las métricas calculadas, la reforma constitucional posee sus comunidades menos separadas que la reforma educacional.
- Componentes Conexas: Se encontraron 628 componentes conexas en la red de la reforma educacional y 549 en la constitucional, esto significa que los usuarios de estas componentes no tienen interacción entre ellos.

  A pesar de que la reforma educacional es una red con menos usuarios que la reforma constitucional posee más componentes, lo que significa que la reforma constitucional posee entre sus componentes, algunas que son individualmente mas grandes que las componentes de la reforma educacional, por ende, la información creada en una componente es capaz de llegar a más usuarios. Dada la conclusión anterior, se debe hacer notar que esta medida no discrimina entre subgrafos mas grandes o más pequeños, por lo que se analizaron los tamaños de las 10 componentes más grandes. Estas componentes son distintas que las comunidades; la comunidad se crea entre usuarios cuya interacción es constante, mientras que la componente conexa considera subgrafos conexos en la red. Por lo tanto, una componente puede tener más de una comunidad, si es que estas poseen un arco que las una.

  La tabla 4.10 muestra que en ambas reformas existe una componente conexa principal y otras pequeñas con muchos menos usuarios. Esto permite a la información tener el potencial de llegar al menos al 86.9 % de usuarios dentro de la componente principal para ambas reformas. Sin embargo, mientras menos arcos existan en la componente principal, menor es la probabilidad de que esta información pueda viajar a través de toda esta cantidad de usuarios. Ésta característica puede ser chequeada por la métrica de densidad

Reforma Educacional			**Reforma Constitucional**		
Componente	**Usuarios**	**Densidad**	**Componente**	**Usuarios**	**Densidad**
1	3327	0.0007	1	4909	0.0006
2	60	0.0333	2	98	0.0204
3	28	0.0714	3	78	0.0266
4	18	0.1111	4	75	0.0266
5	17	0.1176	5	47	0.0425
6	16	0.1250	6	15	0.1333
7	16	0.1250	7	11	0.1818
8	14	0.1648	8	11	0.1818
9	13	0.1538	9	10	0.2000
10	12	0.1818	10	9	0.2222

Tabla 4.10: Componentes más grandes en cada reforma

(Fuente: Elaboración Propia)

- Densidad: Para las redes el valor de la densidad es 0.0006 (R. Educacional) y 0.0005 (R. Constitucional), lo que comprueba que existen pocas conexiones entre los usuarios en comparación al potencial número de arcos, sin embargo, para apoyar la conclusión anterior se calculó la densidad de las diez componentes conexas más grandes para cada reforma (ver tabla 4.10). Se puede ver que los valores de densidad aumentan, sin embargo, dado que la densidad toma valores entre cero y uno, siguen siendo valores bajos, lo que apoya la conclusión anterior sobre la poca probabilidad de difusión de la información, sobre todo en la componente más grande.

- Conjuntos independientes maximales: Utilizando la métrica de conjunto independiente maximal, en promedio para la reforma educacional y constitucional poseen aproximadamente el 93.9 % y 91.4 % de sus nodos sin ninguna relación entre ellos. Esto quiere decir que el 6.1 % y 8.6 % de usuarios restantes de cada red son los que cumplen la función de puentes en la difusión de la información con el resto de usuarios y en consecuencia, son los que principalmente mantienen la red con vida. Este número tan alto muestra que la red no es equitativa dado que la mayor parte del contenido que le da vida a la red debe viajar a través alguno de estos pocos usuarios, y si estos usuarios desaparecen, la red se pierde.

  Es decir, en promedio, el 6.1 % de usuarios mantienen la red viva en la reforma educacional, y el 8.6 % en la reforma constitucional, haciendo a esta última más democrática al tener una proporción mayor de usuarios conectores a pesar de ser una red mucho mas grande. Sin embargo, ambos son valores bajos para hablar de redes equitativas en cuanto a la información creada por cada usuario.

Redes Según Hashtag

Dado el tamaño de las redes y el ruido existente, se analizarán las redes de ambas reformas según el uso de sus hashtags (oficiales y no oficiales), con el objetivo de conocer el comportamiento de los actores para ambos tipos de hashtags, y en cual de estos existe mayor participación, mediante el mismo análisis hecho a las redes generales.

Al observar la grafica de las redes no oficiales (ver figura 4.12 y 4.13) pueden observarse a simple vista, comunidades separadas, donde una de ellas corresponde al ruido de los datos, el cual gracias a la construcción de la red ha logrado aislarse en la gráfica. Las redes oficiales por su parte(ver figura 4.14 y 4.15) pueden verse mas limpias, encontrándose en su mayoría usuarios pertenecientes a Chile. De lo anterior se puede deducir que posterior al filtro de estimación de ubicación aplicado en el análisis de ruido, las redes creadas mediante un hashtag oficial, pueden llegar a ser mas confiables que las redes no oficiales, dado el dinamismo de estos últimos tipos de hashtags y el uso que le dan sus usuarios.

Las tabla 4.11 muestra las métricas anteriormente analizadas, aplicadas esta vez a las redes oficiales y no oficiales, donde las no oficiales resultaron ser mas grandes que sus contrapartes oficiales. Si bien se pueden ver números similares a las redes completas existen algunas diferencias, como los conjuntos independientes, que muestran que los porcentajes de usuarios que dan vida a la red se redujo, volviendo a estas redes menos equitativas que las generales, lo que puede ocurrir por separar las redes perdiendo muchos arcos conectores lo cual reduce la cantidad de usuarios conectores para ambas redes.

La densidad es otra métrica que se ha visto reducida, implicando que existía una comunicación entre las redes oficiales y no oficiales, la cual se perdió al separarlas. Dada esta separación, es posible ver otros actores que anteriormente no eran visibles como CNN y la radio de la Universidad de Chile, siendo estos, usuarios de los hashtags oficiales para las reformas analizadas. Por último, las componentes conexas se han reducido, lo cual es una consecuencia natural de haber dividido la red

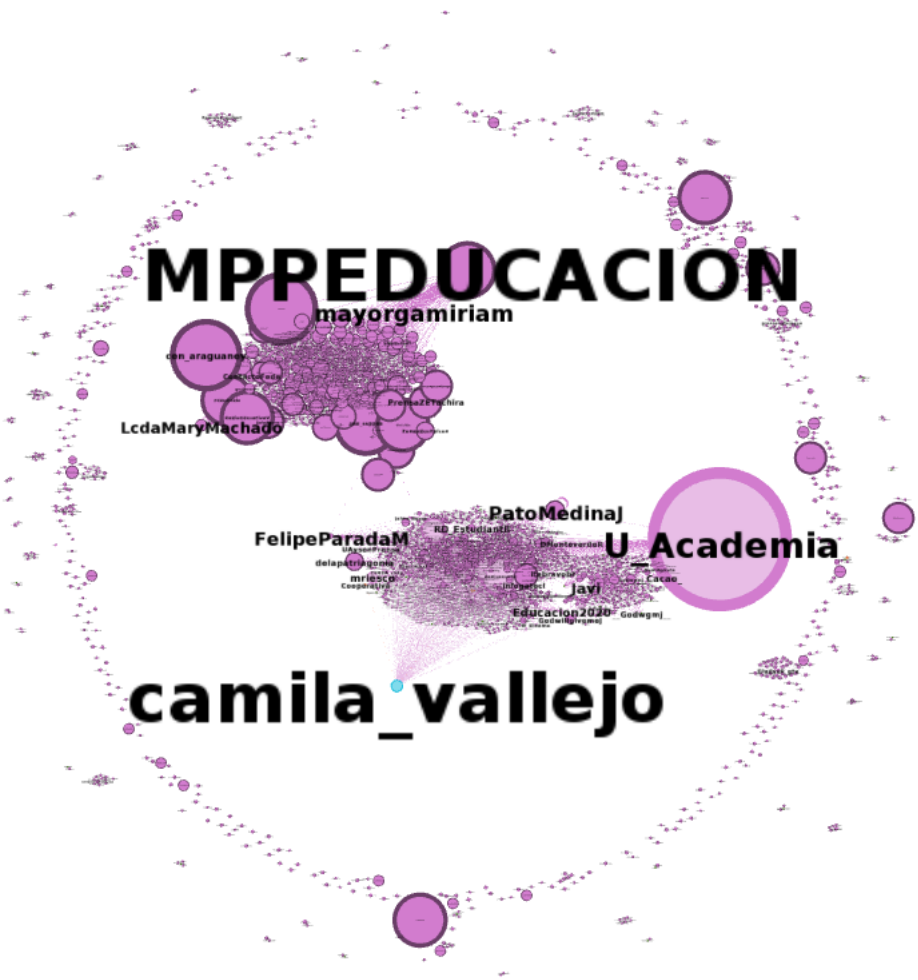


Figura 4.12: Red de Hashtags No Oficiales, Reforma Educacional

Fuente: Elaboración propia

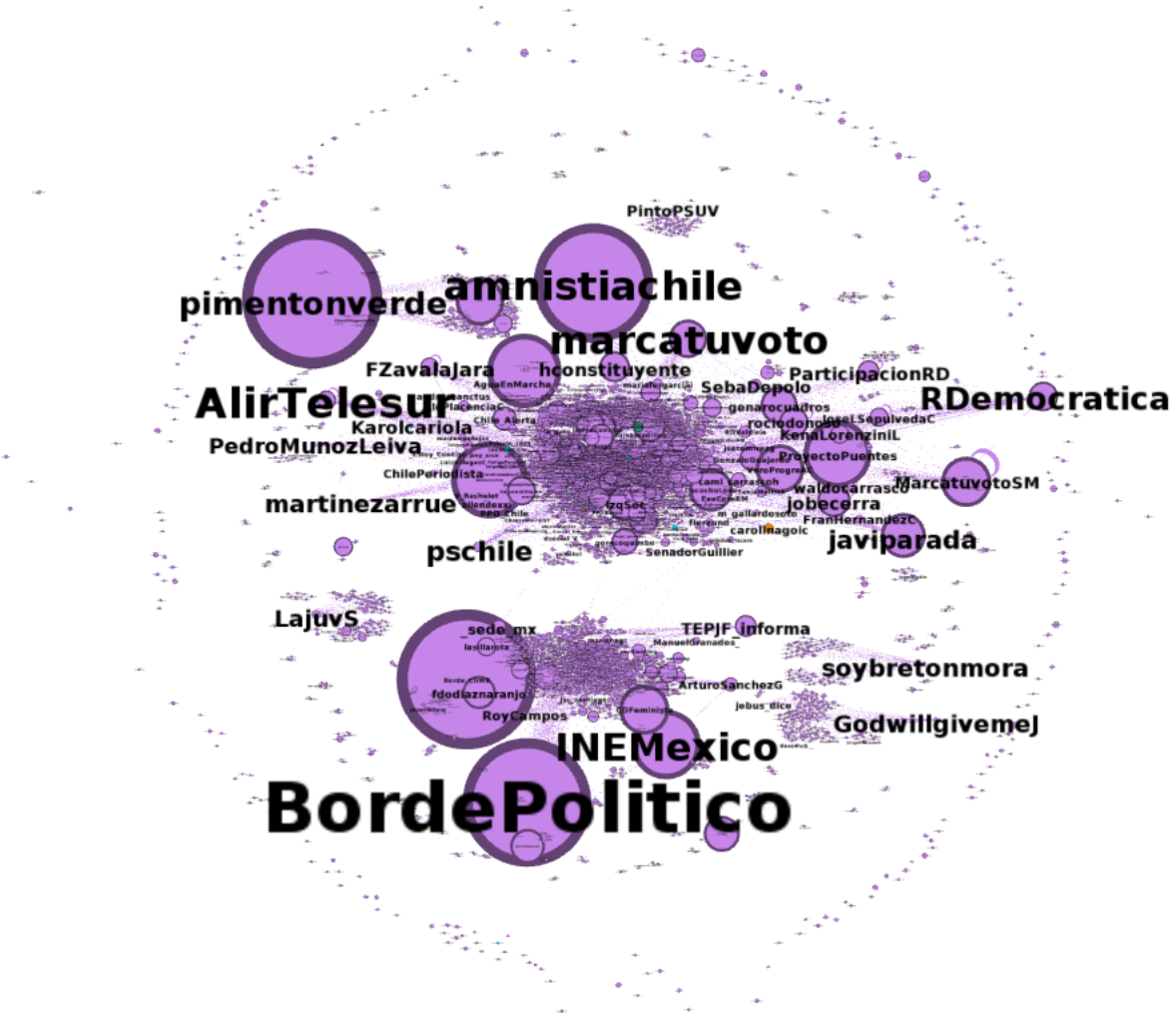


Figura 4.13: Red de Hashtags No Oficiales, Reforma Constitucional

Fuente: Elaboración propia

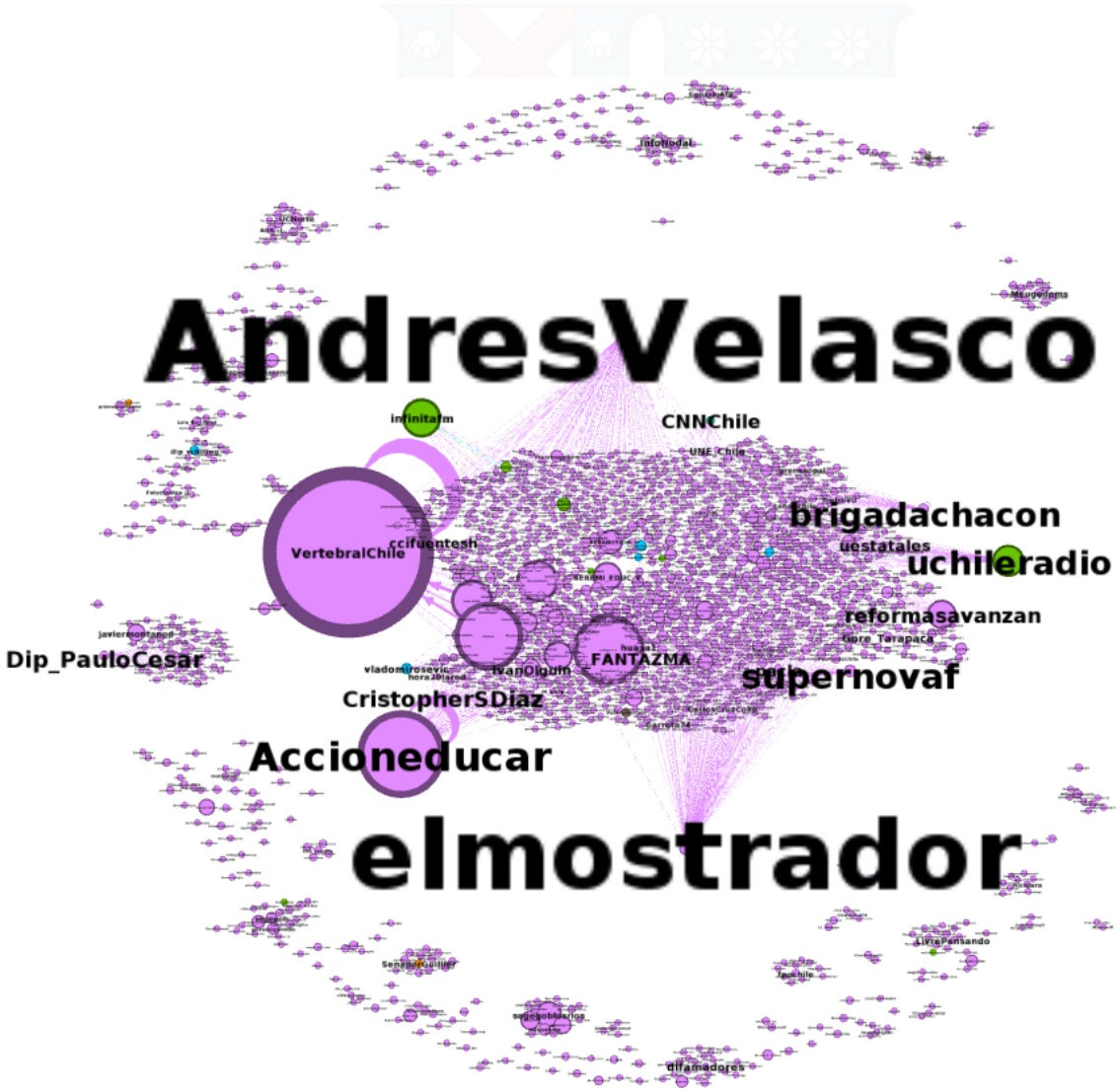


Figura 4.14: Red de Hashtags oficiales, reforma educacional

Fuente: Elaboración propia

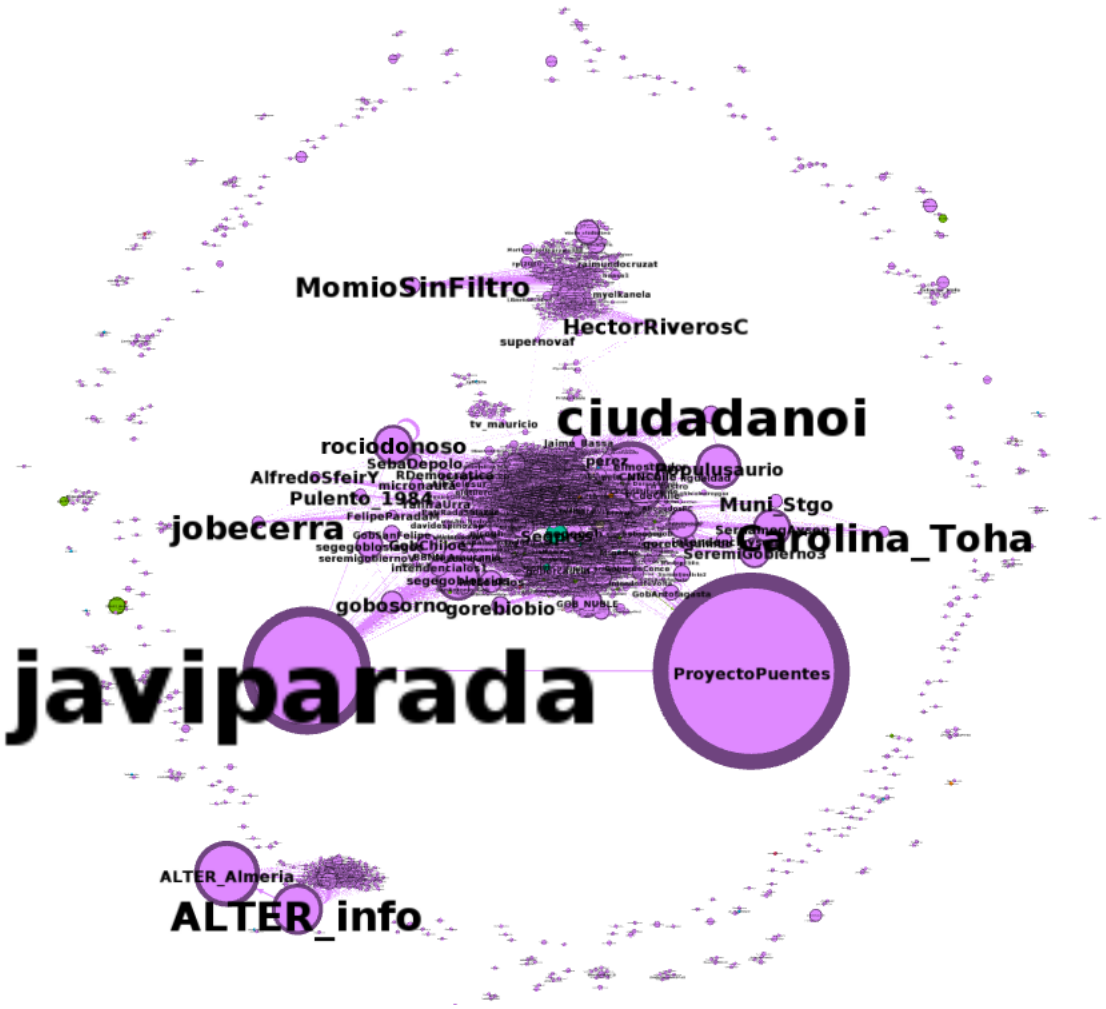


Figura 4.15: Red de Hashtags oficiales, reforma constitucional

Fuente: Elaboración propia

	R. Educacional		**R. Constitucional**	
	H. Oficial	H. No Oficial	H. Oficial	H. No Oficial
Vertices	1586	2957	2974	3877
Arcos	1417	3179	3733	4595
Peso de Arcos				
Media	1.500	1.472	1.493	1.410
Min	1	1	1	1
Max	47	54	15	36
Desviación Estándar	1.973	1.656	1.071	1.027
Mediana	1	1	1	1
Métricas Generales				
Grado	1.784	2.135	2.474	2.348
Coef. Agrupamiento	0.0030	0.0310	0.0440	0.0230
Densidad	0.0005	0.0003	0.0004	0.0003
Componentes Conexas	255	439	280	377
Conj. Independiente Maximal	94.33 %	94.15 %	91.16 %	91.49 %

Tabla 4.11: Reforma Educacional: Tabla resumen según hashtags

(Fuente: Elaboración Propia)

Al igual que en el análisis de las redes completas, se realizaron los T-test para determinar si existen diferencias entre las distribuciones de los grados, esta vez para las reformas oficiales y no oficiales bajo las siguientes hipótesis

- H_0: No existen diferencias en el comportamiento de los grados dependientes del uso de hashtag oficial o no oficial
- H_1: Si existen diferencias en el comportamiento de los grados dependientes del uso de hashtag oficial o no oficial

En este análisis, bajo el valor *p-value*< 0,05, a diferencia del análisis anterior no existe evidencia para rechazar la hipótesis nula (ver tabla 4.12) en ninguno de los dos casos.

	R.E No Oficial	**R.C No Oficial**
R.E Oficial	-1.167 / p = 0,243	
R.C Oficial		0.693 / p = 0,488
$p < 0{,}05$		

Tabla 4.12: T-test en distribución de grados

(Fuente: Elaboración Propia)

Conclusiones Generales

- Cada una de las redes principales posee diferente participación de usuarios. La reforma educacional posee participación mayoritaria de ciudadanos (militantes de partidos), un actor del poder legislativo (Camila Vallejo), y una universidad, mientras que la reforma constitucional posee a simple vista solo presencia ciudadana entre sus usuarios mas participativos. Sin embargo al analizar algunos de estos usuarios, se puede ver que son cuentas de colectivos ciudadanos e iniciativas sociales en vez de ciudadanos individuales, lo cual da mayor representatividad a lo que publican estas cuentas dado que en teoría representan una cantidad mayor de voces en su calidad de organización social.
- Al dividir las redes generales de acuerdo al tipo de hashtag usado puede verse más claramente la participación de algunos medios de prensa y otras entidades políticas, solo en la red oficial (ver figura 4.14), lo que muestra la tendencia de algunos actores de usar hashtags oficiales por sobre los no oficiales. En el caso de los medios de comunicación, esto se debe a mantener la neutralidad en la información entregada, mientras que en el caso de los actores políticos no se tiene suficiente información para obtener una explicación concluyente. Esto se complementa al análisis de uso de hashtags realizado anteriormente, donde las conclusiones para los medios de prensa fueron las mismas.
- Todas las redes comparten características similares en cuanto al peso de sus arcos. Al momento de analizar los grados, cada usuario de dichas redes interactúa con otros dos, en base a las medidas obtenidas en los grados de cada red. Sin embargo, a pesar de ser pocos, existen usuarios pertenecientes al poder legislativo y medios de prensa quienes poseen grados mucho mayores que el promedio de la ciudadanía, probablemente, gracias a su connotación pública de forma offline. Esto responde la pregunta de investigación relacionada al dominio por parte de ciertos actores sociales y personajes influyentes en la red, donde se puede ver que el dominio de la red está determinado principalmente por figuras públicas, personajes vinculados al mundo político (militantes, legisladores, ministros, etc.), colectivos sociales y en menor medida medios de prensa.
- Cada red entre las dos principales posee alrededor de 500 a 600 componentes conexas, todas muy separadas entre ellas, como se puede ver mediante las métricas de coeficiente de agrupamiento y densidad. Para la reforma educacional su coeficiente de agrupamiento es menor para la red oficial, pero su densidad es mayor, en comparación a la red no oficial, lo que significa que los vecinos de cada vértice de la red se encuentran menos comunicados en comparación a la red no oficial, pero a su vez la red completa posee mayor conectividad.

- Las redes no poseen un comportamiento equitativo, dado que la mayor parte de la información circulante en cada una es diseminada gracias a menos del 10 % de usuarios para la reforma educacional y constitucional respectivamente, siendo los valores incluso menores al separar las redes (cercanos al 8 %) Ésta información, generada por los distintos usuarios de la red, debe pasar y ser republicada por esta minoría de usuarios para viajar por la mayor parte de la red, lo cual si bien permite la difusión, da el control de la información a un grupo reducido de usuarios.

Análisis de Influencia

En la sección anterior, se pudo observar en cada red nodos con un tamaño mayor a otros. El tamaño de estos nodos hablan de participación del usuario en la red, Sin embargo, no necesariamente estos usuarios son los más influyentes, dependiendo del enfoque con el que se quiera medir la influencia. En esta investigación se tomó en cuenta dos enfoques relacionados a retweets para compararlos con el análisis de grado de los vértices realizado anteriormente. A éstos enfoques se les llamó influencia temporal e influencia constante. Estos fuerón aplicados a todos los usuarios de manera independiente, obteniendo los diez usuarios más influyentes para cada red y cada definición de influencia, mientras que a cada actor social analizado se le aplicó de manera general un análisis de influencia, para obtener el actor social mas influyente.

Influencia temporal en usuarios

Como puede verse en las tablas 4.13 y 4.14, los usuarios con mas influencia utilizando la métrica de influencia temporal son parte de la ciudadanía, un simple análisis utilizando twitter muestra que la mayor parte son dirigentes, ex dirigentes y militantes de partidos políticos (Principalmente el partido Comunista de Chile), en el caso de la reforma Educacional. Para el caso de la reforma constitucional, la lista está conformada por ciudadanos, sin más información de militancia. Los usuarios cuyo rol no aparece en la tabla, son cuentas que al momento de realizar estos análisis habían sido eliminadas de Twitter sin un motivo conocido.

Usuario	Rol	N° Tweets	N° Retweets	RT promedio por Tweet
jalbertdir	Ciudadano (E)	1	334	334
Mhalberr	Ciudadano	1	333	333
RobertoFavioP	Ciudadano	1	333	333
Tania_Fierro	Ciudadano	1	333	333
rigorh	Ciudadano	2	667	333
1524patricia	Ciudadano	1	328	328
63661f574e4e4ca	Ciudadano	1	328	328
Adolfocabezas1	Ciudadano	1	328	328
AirwalkChile	Ciudadano	1	328	328
Ale_Basulto	Ciudadano	1	328	328
AlonsoFerreiraV	Militante	1	328	328

(E) : Extranjero

Tabla 4.13: Top 10 Influencia Temporal: Reforma Educacional

(Fuente: Elaboración Propia)

Usuario	Rol	N° Tweets	N° Retweets	RT promedio por Tweet
NeftaliOrtizV	Ciudadano (E)	2	544	272
AngelZambranoR3	Ciudadano (E)	2	276	138
DeCondorito	Ciudadano	2	276	138
SSATYR	Ciudadano	2	276	138
mario_ramon	-	2	276	138
xmenina	Ciudadano	2	276	138
AkitoCL	Ciudadano	1	117	117
Alvaror23811980	Ciudadano	1	117	117
Aranguiz79	Ciudadano	1	117	117
BorisSilva1989	Ciudadano	1	117	117

(E) : Extranjero

Tabla 4.14: Top 10 Influencia Temporal: Reforma Constitucional

(Fuente: Elaboración Propia)

Influencia constante en usuarios

Usando la métrica de influencia constante, para las reformas educacional y constitucional, los usuarios influyentes cambian sumándose usuarios pertenecientes al poder legislativo (ver tablas 4.15 y 4.16), como Camila Vallejo, y a otros movimientos sociales e iniciativas, como marcatuvoto y personalidades influyentes en el mundo offline como javiparada, pero sin pertenencia a los actores estudiados.

Usuario	Rol	N° Tweets	N° Retweets	RT promedio por Tweet
U_Academia	Universidad	149	1269	8
luna_rojav	Periodista	28	1151	41
MPPEDUCACION	Ministerio (E)	59	1132	19
anaderodriguez	Ciudadano	18	843	46
danicaucoto	Ciudadano	40	842	21
messinagodoy	Ciudadano	15	813	54
camila_vallejo	Diputada	14	800	57
bespierre	Ciudadano	7	749	107
Gallar13G	Ciudadano (E)	75	748	9
alvaroramis	Escritor	6	738	123

(E) : Extranjero

Tabla 4.15: Top 10 Influencia Constante: Reforma Educacional

(Fuente: Elaboración Propia)

Usuario	Rol	N° Tweets	N° Retweets	RT promedio por Tweet
javiparada	Activista	145	1337	9
roxmapa	Ciudadano (E)	10	892	89
Chile_Alerta	Activista	31	848	27
AlirTelesur	-	24	802	33
TuiterosConMB	Colectivo	79	722	9
felipzuniga	Ciudadano	50	693	13
rociodonoso	Ciudadano	55	654	11
DafneConchaF	Ciudadano	41	653	15
jobecerra	Periodista	34	567	16
marcatuvoto	Colectivo	32	545	17

(E) : Extranjero

Tabla 4.16: Top 10 Influencia Constante: Reforma Constitucional

(Fuente: Elaboración Propia)

Influencia en actores

A partir del análisis anterior se puede ver que de acuerdo a la definición de influencia aplicada pueden aparecer usuarios de distintos actores sociales, por lo cual es importante realizar un análisis general de influencia sobre los cuatro actores estudiados. Para esto se analizará la información de generación de tweets y retweets por cada actor obtenida en el análisis descriptivo, proporcional a la cantidad de usuarios por cada grupo. Se tomará como medida de influencia la cantidad de Retweets por cada Tweet publicado.

La tabla 4.17 muestra que en promedio los actores mas influyentes son la ciudadanía y el poder legislativo, siendo los únicos capaces de multiplicar la información que publican en un factor de 30 veces aproximadamente para la reforma educacional y 15 para la reforma constitucional, en comparación a los otros dos actores, cuyos factores son mucho menores o simplemente no se hayan presentes. Este resultado es importante debido a que la cantidad de usuarios en el poder legislativo es mucho menor que la cantidad de usuarios presentes en el conjunto ciudadanía, por lo que a nivel individual, un usuario del poder legislativo es más influyente que un ciudadano, sin embargo en la red las personas agrupadas como ciudadanía logran influir tanto o más que estos personajes influyentes. Ante estos resultados, no se conoce aún si este comportamiento es igual para todos los usuarios de cada conjunto o existen usuarios que destacan en su propia categoría en cuanto a influencia.

	R. Educacional			**R. Constitucional**		
	Tweets	Retweets	Retweets/Tweet	Tweets	Retweets	Retweets/Tweet
Legislativo	30	888	30	48	698	15
Ejecutivo	0	0	0	24	94	4
Prensa	37	136	4	47	156	3
Ciudadanía	9904	378594	38	16162	220466	14

Tabla 4.17: Influencia promedio por actor social

(Fuente: Elaboración Propia)

Las figuras 4.16 y 4.17, muestran las curvas de tweets, retweets y el cociente entre retweets y tweets. Para ambas reformas se puede ver que el comportamiento de los distintos actores es similar entre ellos al momento de Twittear. Es decir, los usuarios del poder legislativo twittean menos de quince veces en los dos meses de recolección de datos, aproximadamente no más de siete veces al mes. Con esta cantidad de tweets, el usuario con menos retweets del poder legislativo consigue más que los primeros 500 ciudadanos con menos retweets por separado, mientras que el usuario con más retweets, consigue 800 bajo esta conducta de publicación, consiguiendo el usuario mas influyente 57 retweets por publicación. Por su parte, los usuarios de la ciudadanía poseen una conducta de publicación mucho más masiva llegando a un máximo de 149 tweets, 1269 retweets y 334 retweets por publicación. Sin embargo, estos son casos de influencia temporal, como se vió en el análisis de usuarios independientes.

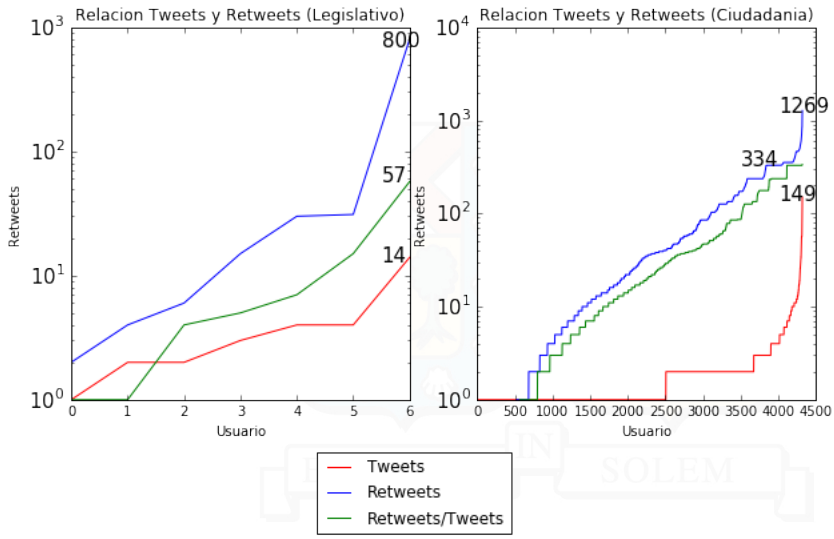


Figura 4.16: Comparativa de influencia, Reforma Educacional

Fuente: Elaboración propia

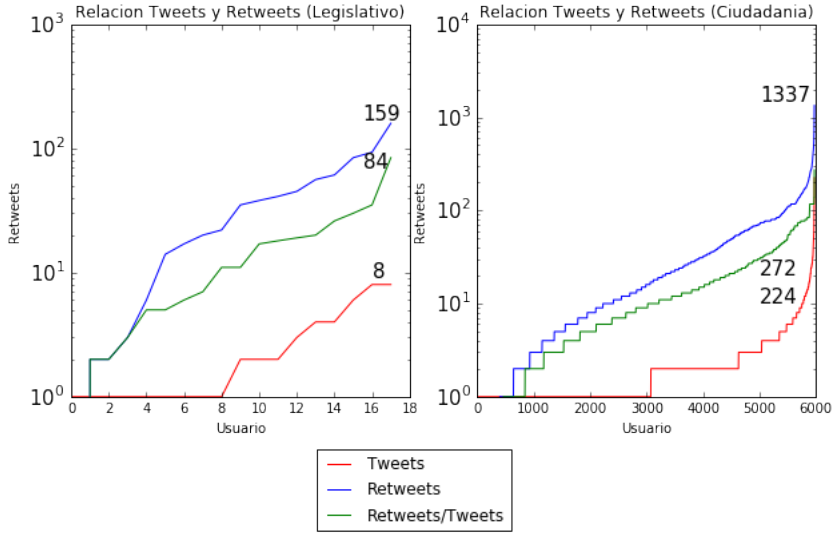


Figura 4.17: Comparativa de influencia, Reforma Constitucional

Fuente: Elaboración propia

Conclusiones Generales

En base a las dos definiciones de influencia, bajo el contexto utilizado se elige la influencia constante como la métrica para esta memoria, debido a que la influencia temporal no es una buena métrica para el corto periodo de recolección de los datos, dado que no se conoce si los usuarios mas influyentes bajo este concepto poseen una baja frecuencia de publicación con una alta cantidad de retweets (para conocer esto habría que analizar la conducta de dichos usuarios en un periodo más largo) o si fue el concepto de "minuto de fama" lo que les permitió tener un impacto temporal en la red.

Para esta definición de influencia, se consiguió determinar que individualmente los usuarios más influyentes si bien existen, no pertenecen mayoritariamente a alguno de los actores sociales externos a la ciudadanía, dejando espacio para que usuarios sin visibilidad mediática, entidades académicas y organizaciones sociales puedan darse a conocer en la red, participar e influir en la agenda pública. Sin embargo, al momento de analizar la influencia de los actores sociales en general, la ciudadanía debe realizar un esfuerzo mucho mayor en cuanto a publicaciones y número de usuarios para poseer un grado de influencia en Twitter comparable al poder legislativo, el cual con una baja cantidad de usuarios y de publicaciones consigue una difusión similar a toda la ciudadanía. Esto no solo se debe al factor mediatico debido a que los medios de prensa y el poder ejecutivo no logra estos niveles de influencia en Twitter. Dichos factores no pueden ser obtenidos mediante los datos actuales.

Análisis de Sentimientos

El objetivo de realizar este análisis es, luego de tener el análisis descriptivo y el análisis de redes, conocer cual es la apreciación de los usuarios sobre las reformas y cuales de las opiniones tiene mayor influencia en la red. Es aquí donde el análisis de contenido juega un papel importante, dado que el análisis anterior no nos dice en ningún caso si los usuarios publican con una connotación positiva o negativa acerca de las reformas y si la información propagada en la red es también positiva o negativa.

Clasificando cada tweet recolectado, por cada reforma, se obtienen los resultados que muestra la tabla 4.18. En esta se puede ver que existe una opinión muy pareja en la reforma educacional, en cuanto a sentimientos positivos y negativos, a diferencia de la reforma constitucional la cual posee casi en su mayoría opiniones positivas. Estos datos muestran que existe un nivel de debate en la primera reforma, la cual presenta una diferencia de ideas mas cercana, mientras que en la segunda reforma la aceptación es mucho mayor.

Actor	Emoción Positiva	Emoción Negativa
Reforma Educacional		
Legislativo	0.19 %	0.11 %
Ejecutivo	0.00 %	0.00 %
Prensa	0.22 %	0.15 %
Ciudadanía	50.4 %	48.64 %
Reforma Constitucional		
Legislativo	0.25 %	0.036 %
Ejecutivo	0.12 %	0.018 %
Prensa	0.21 %	0.067 %
Ciudadanía	79.7 %	19.1 %

Tabla 4.18: Análisis de sentimientos: Tabla resúmen

(Fuente: Elaboración Propia)

Conociendo las proporciones de tweets positivos y negativos dentro de los datos, se realizó una clasificación en base a la cantidad de retweets por cada emoción. La figura 4.18 muestra que para la reforma educacional si bien proporcionalmente hay igual cantidad de tweets negativos y positivos, se ve que las publicaciones de connotación negativa poseen una mayor cantidad de retweets, mientras que para la reforma constitucional la mayor parte de sus publicaciones retwitteadas son de connotación positiva. Esto da fuerza a la sospecha obtenida en el análisis de hashtags donde se concluye que la reforma Educacional es un tópico mucho mas conflictivo para los usuarios de Twitter, lo cual explica la movilización social en el contexto *offline* dado que la información de connotación negativa posee más influencia en la red y en sus usuarios. En el caso de la reforma constitucional, la información positiva es la que influencia la red, lo que puede implicar un posible apoyo de los usuarios de Twitter a esta reforma o al cambio constitucional en si mismo.

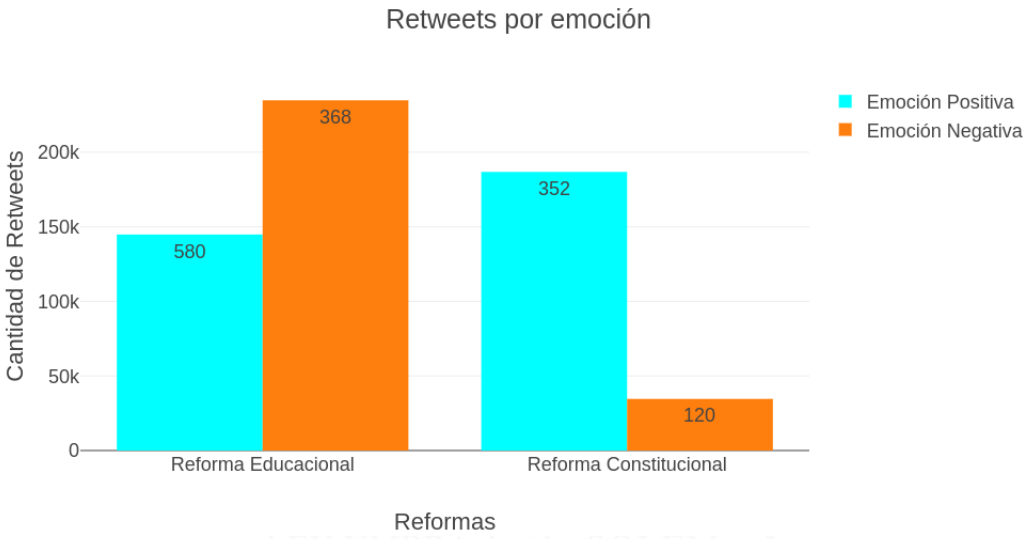


Figura 4.18: Cantidad de Retweets por emoción

Fuente: Elaboración propia

Por último, se analizaron los 10 usuarios mas influyentes según influencia constante, debido a que son los que poseen mas retweets, con el objetivo de conocer si entre los usuarios mas influyentes se difunde por la red más información positiva o negativa. Según la tabla 4.19, el comportamiento de los usuarios más influyentes es representativo del comportamiento general de todos los usuarios. La proporción entre tweets positivos y negativos no es tan diferente, superando por poco los tweets positivos, pero al momento de analizar los retweets totales se ve que la mayor parte del contenido retwitteado corresponde a publicaciones de connotación negativa. Para el caso de la reforma constitucional la mayor parte de los tweets son positivos, al igual que los retweets siendo sus usuarios más influyentes difusores de contenido positivo a diferencia de la reforma anterior.

Usuario	**Tweets Pos.**	**Tweets Neg.**	**Retweets Pos.**	**Retweets Neg.**
U_Academia	115	34	661	608
luna_rojav	19	9	510	641
MPPEDUCACION	9	50	165	967
anaderodriguez	9	9	539	304
danicaucoto	21	19	309	533
messinagodoy	10	5	285	528
camila_vallejo	9	5	272	528
bespierre	4	3	69	680
Gallar13G	35	40	306	442
alvaroramis	3	3	58	680
Total	234	177	3174	5911

Tabla 4.19: Influencia de sentimientos en reforma educacional

(Fuente: Elaboración Propia)

Usuario	Tweets Pos.	Tweets Neg.	Retweets Pos.	Retweets Neg.
javiparada	108	37	1050	287
roxmapa	10	0	892	0
Chile_Alerta	22	9	712	136
AlirTelesur	15	9	695	107
TuiterosConMB	72	7	610	112
felipzuniga	47	3	678	15
rociodonoso	48	7	586	68
DafneConchaF	32	9	456	197
jobecerra	30	4	458	109
marcatuvoto	28	4	495	50
Total	412	89	6632	1081

Tabla 4.20: Influencia de sentimientos en la reforma constitucional

(Fuente: Elaboración Propia)

Conclusiones generales

El contenido generado por los usuarios de la reforma educacional, si bien es equilibrado entre positivo y negativo, existe una tendencia en la red a retwittear las publicaciones negativas, lo que permite concluir que existe un rechazo ante el proyecto de reforma educacional por parte de la red analizada, a diferencia de la reforma constitucional donde se genera mayoritariamente contenido positivo y este mismo es el que se retwittea en mayor parte. Este análisis, en complemento al análisis de redes permite intuir cuando la situación social respecto a una ley es más delicada o menos aprobada por un segmento de la ciudadanía.

Validación de análisis

Análisis de Sentimientos

Dado que para el análisis de sentimientos se utilizaron datos de España, los cuales si bien están en el idioma español, poseen un contexto social diferente al chileno. Por lo tanto, para clasificar datos nuevos se debe realizar un análisis cualitativo y cuantitativo para probar la efectividad del método utilizado.

Para validar el modelo se creó un dataset de testing utilizando el 10 % de tweets de cada reforma analizada (Educacional y Constitucional), teniendo un total de 997 y 1328 tweets respectivamente para cada reforma. Cada uno de estos tweets fue clasificado de manera manual por dos personas diferentes, para luego ser clasificados utilizando el modelo creado, y comparar la precisión de este para cada conjunto de datos. Estos resultados se ven en la tabla 5.1, donde considerando que el modelo posee una precisión de testing del 80.4 % para la data original, para la reforma Constitucional la precisión no varía demasiado, consiguiendo un buen valor.

En el caso de la reforma educacional, la precisión posee una baja importante dejándola en aproximadamente un 50 %, lo cual puede deberse a diferentes usos de vocabulario, por poseer un contexto social muy diferente de los datos originales o por su mayor porcentaje de ruido debido a la data extranjera. Dados estos resultados la validación del método se realizará utilizando la reforma constitucional, debido a que es la reforma con más datos, más limpia en cuanto a ruido y con mayor precisión respecto a la anotación manual.

Dataset	Anotador 1	Anotador 2
Ref. Educacional	49.05 %	53.66 %
Ref. Constitucional	75.68 %	78.32 %

Tabla 5.1: Validación de Análisis de sentimientos: Accuracy

(Fuente: Elaboración Propia)

Las diferencias en la clasificación manual entre cada anotador y el clasificador pueden ser vistas en la tabla 5.2, donde se puede ver a priori que existen pocas diferencias en entre los anotadores y el clasificador. Para comprobar esto, se calculó el **coeficiente Kappa de Cohen**.

Reforma Constitucional	**Anotador 1**	**Anotador 2**	**Clasificador**
Positivo	77.33 %	87.10 %	83.35 %
Negativo	22.67 %	12.90 %	16.65 %

Tabla 5.2: Validación de Análisis de sentimientos: Etiquetado

(Fuente: Elaboración Propia)

Para el conjunto de datos de la reforma constitucional, se calculó el coeficiente Kappa de Cohen, *accuracy* y *recall* para ambos anotadores (ver tabla 5.3), junto con la matriz de confusión para cada anotador (ver tablas 5.4 y 5.5).

El valor del Kappa de Cohen si bien no es un acuerdo perfecto, se encuentra entre un acuerdo razonable y moderado, ubicándose en el centro de la tabla 3.5, lo cual significa que el proceso de anotación posee confiabilidad, aunque no perfecta.

Por su parte, los valores de *accuracy* y *recall* son altos para cada uno de los anotadores, lo que le entrega confiabilidad al proceso de análisis de sentimientos, respaldado por las matrices de confusión para cada anotador, las cuales muestran valores altos en los verdaderos positivos y verdaderos negativos, los cuales son los valores concordantes entre el valor real y el predicho.

	Accuracy	**Recall**
Anotador 1	75.68 %	88.17 %
Anotador 2	78.32 %	85.40 %
Kappa de Cohen	0.39	

Tabla 5.3: Validación de Análisis de sentimientos: Etiquetado

(Fuente: Elaboración Propia)

Conclusiones Generales

Como conclusión de esta validación se obtiene que si bien el método mantiene una tasa de confiabilidad y representatividad al momento de clasificar la información, está lejos de ser perfecto debido tanto a la precisión de clasificación del modelo, como también al ruido en los datos, por lo cual una alternativa a futuro sería analizar otro tipo de modelos utilizando otras técnicas, como por ejemplo, redes neuronales y someter a los datos a filtros y procesamientos que permitan que esta llegue más limpia al clasificador.

		Valor predicho		
		p	**n**	**total**
Valor real	**p′**	True Positive: 1110	False Negative: 149	P′: 1259
	n′	False Positive: 247	True Negative: 122	N′: 369
	total	P: 1357	N: 271	

Tabla 5.4: Matriz de confusión: Anotador 1

(Fuente: Elaboración Propia)

		Valor predicho		
		p	**n**	**total**
Valor real	**p′**	True Positive: 1211	False Negative: 207	P′: 1418
	n′	False Positive: 146	True Negative: 64	N′: 210
	total	P: 1357	N: 271	

Tabla 5.5: Matriz de confusión: Anotador 2

(Fuente: Elaboración Propia)

EX UMBRA
IN
SOLEM

Conclusiones

Sobre la investigación realizada

A partir del trabajo desarrollado en esta memoria se ha logrado crear una metodología que mediante un análisis descriptivo, SNA y análisis de sentimientos se puede conocer el comportamiento de usuarios en Twitter a través de un conjunto de datos recolectado previamente. Estos análisis permitieron conocer el modelo de influencia y las conductas de uso de los distintos actores sociales estudiados tanto de manera individual, como su aporte en la totalidad de la red. Estas redes tienen todas la característica de ser poco conectadas, donde los usuarios en general no interactúan constantemente entre ellos. Además se encuentran segmentadas en pocos clusters de gran tamaño y muchos clusters pequeños los cuales no se comunican entre si y dependientes de una minoría de usuarios que mantienen la red viva.

Antes de realizar esta investigación, fueron planteadas una serie de preguntas a las cuales estos análisis fueron enfocados para responder, consiguiendo para cada una de estas, respuestas respaldadas por dichos análisis y sus replicas.

Respuesta a preguntas de investigación

- **¿Están las redes sociales realmente dando libertad y visibilidad de las opiniones de cada usuario por igual?**

  Al analizar las frecuencias de publicación y difusión de los tweets se observó que para cada usuario dichas frecuencias no son las mismas, en efecto, existen usuarios los cuales poseen mas retweets que el promedio. Sin embargo, al analizar esta distribución de retweets y sometiéndola al coeficiente de Gini, se pudo concluir que existe una alta desigualdad en la difusión del contenido creado por unos usuarios por sobre otros, derrumbando en este caso la idea de que Twitter crea redes democráticas.

Si bien podría ser normal pensar que existen usuarios que por poseer mayor visibilidad mediática tengan mayor difusión de su contenido, el coeficiente de Gini es lo suficientemente alto para afirmar que en este caso las redes no son democráticas. Esto es apoyado al analizar las métricas de las redes donde aproximadamente el 10 % de los usuarios actúan de conectores para difundir la información a lo largo de la red, dado que sin ellos, la red se pierde. Es decir, la difusión exitosa de la información que publica cualquier usuario en esta red depende en si estos usuarios conectores republican dicho contenido, dándoles el control sobre la información que publica el 90 % de la red.

Esto demuestra que en la red existen usuarios que impactan más la red por sobre otros, es decir, usuarios influyentes que determinan lo que se discute en Twitter.

- **¿Quienes son los usuarios influyentes en una red de usuarios en un contexto determinado?**

  En esta memoria se trabajó con datos de Twitter en un contexto político, analizando los tweets que hablaron sobre las principales reformas del segundo mandato de la presidenta Michelle Bachelet. Dado este contexto, la búsqueda de usuarios influyentes se realizó tomando en cuenta las métricas de grado de las redes y el concepto de influencia constante. Ambas pruebas arrojaron resultados similares e incluso un conjunto similar de usuarios influyentes. Una conclusión interesante es que los usuarios del poder legislativo, ejecutivo y la prensa no se encuentran de manera abundante entre los usuarios mas influyentes, observándose solo un usuario del poder legislativo y la mayor parte de usuarios en la categoría de ciudadanos, encontrándose entre estos militantes de partidos políticos, otros personajes influyentes en el mundo *offline* y cuentas de colectivos e iniciativas sociales. La presencia de estos últimos permite concluir que en Twitter, los personajes y organizaciones que actúan como lideres de opinión son diferentes a los que se encuentran en el modelo offline, otorgando valor como plataforma alternativa a los medios de comunicación tradicionales, dado que entrega espacio a organizaciones sociales y conjunto de ciudadanos para comunicarse con la ciudadanía y poner en la agenda pública temas relevantes y diferentes a los ofrecidos por los medios tradicionales.

- **¿Existe dominio por parte de ciertos actores sociales?**

  Si bien los usuarios influyentes corresponden a distintos actores sociales, la situación cambia al analizar cada actor social como un conjunto. Analizando el comportamiento de difusión y emisión de tweets, junto a la influencia por actores se observó que los actores mas influyentes en la red son la ciudadanía y el poder legislativo, cuya información es la que mas se difunde en la red. Sin embargo, esto implica que el poder legislativo con sus 151 usuarios logra igualar a la ciudadanía con 12729 usuarios, por lo tanto el actor social mas influyente proporcional a su número de usuarios es el poder legislativo, seguido de la ciudadanía. En cambio, el poder ejecutivo posee poca o nula presencia en lo que se discute de las reformas, mientras que la prensa posee cobertura trascendental, pero en baja medida.

- **¿Existe un dominio de la prensa en la información online?**

  Dada la pregunta de investigación anterior, se concluye que los medios de prensa en Twitter cumplen su función de medio de comunicación dentro de la red consiguiendo posicionarse entre los usuarios de esta. Sin embargo, no poseen un control de la información que se difunde en la red, dado que no poseen un nivel de influencia elevado. Estos medios se caracterizan por utilizar principalmente los hashtags oficiales y no hacerse parte de las discusiones en Twitter, funcionando sólo como distribuidores de contenido, comprobando así que no se replica el modelo *offline* en Twitter, el cual posee su propia dinámica de flujo de información con otros actores sociales y actores que controlan e influyen en los temas discutidos.

- **¿Que clase de información es la que se difunde a través de la red?**

  Por último, utilizando el análisis de sentimientos se pudo concluir que la proporción de publicaciones positivas y negativas en la reforma educacional es similar, pero existe una tendencia de los usuarios de la red a compartir principalmente las publicaciones negativas, mientras que en la reforma constitucional prima el contenido de connotación positiva y de igual forma es el que mas se difunde en la red. Este análisis es un complemento importante a SNA, dado que permite conocer el tipo de información que circula por la red, concluyendo así que la reforma educacional es un tema de conflicto en Twitter donde se puede deducir que los usuarios están en contra en una mayor parte, a diferencia de la reforma constitucional que se puede encontrar a lo largo de la red como comentarios positivos.

Factores de complicación del análisis

En la ejecución de esta metodología en las reformas seleccionadas se encontraron distintos factores que deben ser tratados cuidadosamente para no tener errores en los datos recolectados y en consecuencia, en los análisis realizados, como para poder realizar estos de manera optima.

En lo que respecta al error en los datos, se encontró que los usuarios no poseen sus localizaciones al momento de publicar, lo que puede causar ruido en la data al obtener tweets de usuarios que usen el mismo hashtag para otro contexto en otro lugar del mundo. Para solucionar estos problemas no solo es necesario realizar una elección correcta y fundamentada de hashtags, sino que también es importante realizar una búsqueda de técnicas de filtrado de tweets para minimizar el ruido en el conjunto de datos. Respecto a la data analizada, se concluye que dado a los riesgos anteriormente mencionados, las redes pertenecientes a la reforma laboral y tributaria debieron ser excluidas de los análisis de redes y de sentimientos debido a la alta fuente de error encontradas en datos pertenecientes a otros países y contextos sociales. Sin embargo, la reforma educacional y constitucional pudieron ser analizadas sin problemas llegando a aíslarse visualmente estas comunidades extranjeras al observar las redes generadas.

Cumplimiento de objetivos

Objetivo General

El objetivo general que se propuso para esta memoria fue: **"Combinar técnicas de recolección y análisis de datos automatizada que permita evaluar si Twitter, como medio de comunicación social y abierto, conserva el modelo de influencia unidireccional desde la prensa a la ciudadanía, o ha dado paso a cambiar dicho modelo."**

Mediante la tecnología de streaming y gracias a la API de Twitter se consiguió ejecutar un procedimiento automatizado de recolección de datos, con los cuales mediante análisis descriptivo, de redes sociales y de sentimientos permitió describir el modelo de influencia y de flujo de información en Twitter dilucidando que dicho modelo es diferente al unidireccional establecido en el contexto offline, encontrándose regido bajo un modelo dinámico en el cual la información se genera desde diferentes fuentes y es controlado por usuarios pertenecientes a diferentes actores sociales.

Basado en esta conclusión, se considera el objetivo general como cumplido dado que se logró describir el modelo de influencia de las redes creadas gracias a los datos de Twitter.

Objetivos Específicos

- **Definir un método de recolección de datos relevantes a la agenda pública y al tema social escogido.**

  Se consiguió implementar un procedimiento de ingesta de tweets a partir de este trabajo el cual es capáz de filtrar datos de acuerdo a usuarios y hashtags que se desean obtener. En base a esta contribución, estudiantes de doctorado de la universidad Técnica Federico Santa María han tomado este algoritmo y ya han construido extensiones que permiten seguir escuchando Twitter para futuras investigaciones. Mientras que los códigos relacionados al desarrollo de esta metodología se han publicado mediante la plataforma Github[18], para libre uso de otros usuarios interesados en realizar SNA.

- **Validar la metodología creada aplicada al caso de las cuatro grandes reformas chilenas del gobierno de Michelle Bachelet.**

  Se lograron responder todas las preguntas de investigación planteadas describiendo el contexto social de los usuarios en Twitter y sus conductas de uso de la plataforma para cada uno de los actores sociales, y como estas conductas definen lo que se discute en la agenda pública de Twitter. Sin embargo, debido al ruido no todas las reformas pudieron someterse a los análisis debido al ruido en los datos.

[18] https://github.com/vansimonsen/twitterProject

- **Proponer una metodología de análisis de influencias de distintos actores sociales en temas de interés público en Twitter**

  Se consiguió crear una metodología de estudio de datos políticos que abarca los ejes de:

  1. Frecuencias de publicación, dominio por parte de los diferentes actores, popularidad de hashtags y equidad en la distribución de publicaciones
  2. Análisis de sentimientos para el contenido de las publicaciones y que clase de información en base a su sentimiento es la que viaja por la red.
  3. *Social Network Analysis* para el comportamiento de los usuarios que publican sobre los tópicos en estudio, definiendo influencias y conectividad de los usuarios que forman parte de la red.

Trabajo Futuro

Durante el desarrollo de esta memoria, se logró recopilar una serie de puntos interesantes para mejorar la metodología.

- Previa investigación para filtrar tweets por localización geográfica: Ésto permitirá tener datos más limpios y mas representativos del estrato social estudiado.
- Mejora en el modelo de análisis de sentimientos: Si bien el clasificador utilizado es confiable, para datos nuevos pierde precisión, por lo cual las opciones son aumentar la data de entrenamiento utilizando tweets de un contexto social más cercano (el chileno en éste caso), y/o cambiar el clasificador explorando otras opciones como son las redes neuronales.
- Redes dinámicas: Poseer la evolución de las redes en el tiempo permite responder otras preguntas de investigación y conocer aún más el comportamiento del segmento social investigado.
- Arquitectura de Big Data: La mayoría de los puntos antes mencionados requieren una mayor cantidad de datos para el análisis, sobre todo porque con más datos, las conclusiones obtenidas son más representativas respecto al entorno investigado (siempre y cuando estén libres de ruido). Para esto, se requiere poder de procesamiento para lograr ejecutar los métodos desarrollados en esta memoria sobre una gran cantidad de datos, por lo tanto, se requiere una arquitectura de big data, como por ejemplo Apache Spark o Hadoop.

EX UMBRA
IN
SOLEM

Bibliografía

Bennett, W Lance y Iyengar, Shanto (2010). The shifting foundations of political communication: Responding to a defense of the media effects paradigm. *Journal of Communication*, 60(1), 35–39. 2.3

Bruns, Axel y Burgess, Jean E (2011). The use of twitter hashtags in the formation of ad hoc publics. In *Proceedings of the 6th European Consortium for Political Research (ECPR) General Conference 2011*. 2

Cárdenas Neira, Camila (2014). Representación de la acción política de los estudiantes chilenos: movilización de significados en redes sociales. *Última década*, 22(40), 57–84. 1

Carrington, Peter J; Scott, John; y Wasserman, Stanley (2005). *Models and methods in social network analysis*, volume 28. Cambridge university press. 2.3.1.1

Cristianini, Nello y Shawe-Taylor, John (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press. 3.2

Dos Santos, Cícero Nogueira y Gatti, Maira (2014). Deep convolutional neural networks for sentiment analysis of short texts. In *COLING* (pp. 69–78). 2.3.1.2

Ediger, David; Jiang, Karl; Riedy, Jason; Bader, David A; y Corley, Courtney (2010). Massive social network analysis: Mining twitter for social good. In *Parallel Processing (ICPP), 2010 39th International Conference on* (pp. 583–593).: IEEE. 2.3.1.2

Feldman, Ronen (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4), 82–89. 2.3.1.2

Fruchterman, Thomas MJ y Reingold, Edward M (1991). Graph drawing by force-directed placement. *Software: Practice and experience*, 21(11), 1129–1164. 3.3.3.2

García, Cristobal; Chauveau, Paul; Ledezma, Javier; y Pinto, Maria (2013). What can social media teach us about protests? analyzing the chilean 2011-12 student movement's network evolution through twitter data. *arXiv preprint arXiv:1308.2451*. 1, 2.3.3.1

GfK Adimark, Pontificia Universidad Católica de Chile (2016). *Encuesta Nacional Bicentenario*. Adimark. 1.1, 2.2

Harrington, Stephen; Highfield, Tim; y Bruns, Axel (2013). More than a backchannel: Twitter and television. *Participations*, 10(1), 405–409. 1.1

Hong, Sounman (2012). Online news on twitter: Newspapers' social media adoption and their online readership. *Information Economics and Policy*, 24(1), 69–74. 1

Hu, Xia; Tang, Jiliang; Gao, Huiji; y Liu, Huan (2013). Unsupervised sentiment analysis with emotional signals. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 607–618).: ACM. 2.3.1.2

Huang, Carol (2011). Facebook and twitter key to arab spring uprisings: report. In *The National*, volume 6. 2.3.3

Kouloumpis, Efthymios; Wilson, Theresa; y Moore, Johanna D (2011). Twitter sentiment analysis: The good the bad and the omg! *Icwsm*, 11(538-541), 164. 2.3.1.1

Kulshrestha, Juhi; Kooti, Farshad; Nikravesh, Ashkan; y Gummadi, P Krishna (2012). Geographic dissection of the twitter network. In *ICWSM*. 2.3

Landis, J Richard y Koch, Gary G (1977). The measurement of observer agreement for categorical data. *biometrics*, (pp. 159–174). 3.3.3.3

McCombs, Maxwell E y Shaw, Donald L (1972). The agenda-setting function of mass media. *Public opinion quarterly*, 36(2), 176–187. 2.3

McQuail, Denis (1994). *Mass communication*. Wiley Online Library. 1.2

Meraz, Sharon (2009a). Is there an elite hold? traditional media to social media agenda setting influence in blog networks. *Journal of Computer-Mediated Communication*, 14(3), 682–707. 2.3

Meraz, Sharon (2009b). Is there an elite hold? traditional media to social media agenda setting influence in blog networks. (pp. 682–707). 2.3.2

Mullen, Tony y Collier, Nigel (2004). Sentiment analysis using support vector machines with diverse information sources. In *EMNLP*, volume 4 (pp. 412–418). 2.3.1.2, 3.2

Pak, Alexander y Paroubek, Patrick (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10. 2.3.1.2

Roth, Dan (1998). Learning to resolve natural language ambiguities: A unified approach. In *AAAI/IAAI* (pp. 806–813). 2.3.1.2

Scheufele, Dietram A. y Tewksbury, David (2007). Framing, agenda setting, and priming: The evolution of three media effects models. *Journal of Communication*, 57(1), 9–20. 2.3

Tarjan, Robert Endre y Trojanowski, Anthony E (1977). Finding a maximum independent set. *SIAM Journal on Computing*, 6(3), 537–546. 3.1

Tumasjan, Andranik; Sprenger, Timm Oliver; Sandner, Philipp G; y Welpe, Isabell M (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10, 178–185. 2.3

Un-Habitat (2008). *State of the World's Cities 2008-2009: Harmonious Cities*. Earthscan. 4.1.5

Valenzuela, Sebastián; Arriagada, Arturo; y Scherman, Andrés (2012). The social media basis of youth protest behavior: The case of chile. *Journal of Communication*, 62(2), 299–314. 2.3.3.1

Vallina-Rodriguez, Narseo; Scellato, Salvatore; Haddadi, Hamed; Forsell, Carl; Crowcroft, Jon; y Mascolo, Cecilia (2012). Los twindignados: The rise of the indignados movement on twitter. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)* (pp. 496–501).: IEEE. 2.3.3

Wasserman, Stanley y Faust, Katherine (1994). *Social network analysis: Methods and applications*, volume 8. Cambridge university press. 2.3.1.1

Watts, Duncan J y Strogatz, Steven H (1998). Collective dynamics of 'small-world'networks. *nature*, 393(6684), 440–442. 3.1

Wersig, Gernot y Neveling, Ulrich (1975). The phenomena of interest to information science. *The information scientist*, 9(4), 127–140. 2.1

Wilson, Theresa; Wiebe, Janyce; y Hoffmann, Paul (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05 (pp. 347–354). Stroudsburg, PA, USA: Association for Computational Linguistics. 2.3.1.2

Wood, James M (2007). Understanding and computing cohen's kappa: A tutorial. *WebPsychEmpiricist. Web Journal at http://wpe. info/*. 3.3.3.3

Wu, Shaomei; Hofman, Jake M; Mason, Winter A; y Watts, Duncan J (2011). Who says what to whom on twitter. In *Proceedings of the 20th international conference on World wide web* (pp. 705–714).: ACM. 2.3

Yitzhaki, Shlomo (1979). Relative deprivation and the gini coefficient. *The quarterly journal of economics*, (pp. 321–324). 4

Zhao, Wayne Xin; Jiang, Jing; Weng, Jianshu; He, Jing; Lim, Ee-peng; Yan, Hongfei; y Li, Xiaoming (2011). *Comparing twitter and traditional media using topic models*. Springer. 2.3.2