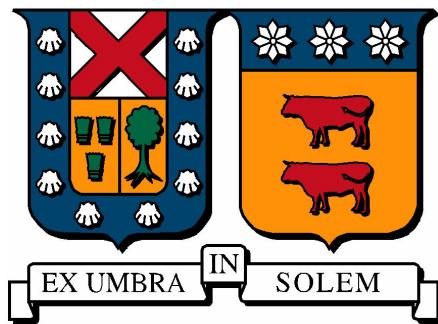


UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INFORMÁTICA
VALPARAÍSO, CHILE



**MODELO PROACTIVO DE DISEMINACIÓN DE
INFORMACIÓN EN REDES MANET ALTAMENTE
DINÁMICAS Y DENSAS**

FRANCO DANIEL CASTRO NAEVA
TESIS PARA OPTAR AL GRADO ACADÉMICO DE
INGERIERO CIVIL INFORMÁTICO

COMISIÓN EVALUADORA:
PROFESOR GUÍA: LAUTARO GUERRA J.
PROFESOR CORREFERENTE:
PROFESOR CORREFERENTE EXTERNO:

NOVIEMBRE 2015

A mi familia, Mario, Yolanda, Mario, Alex y Barrabas.

Índice general

Índice general	III
Índice de figuras	VI
Índice Algoritmos	VII
1. Introducción	1
1.1. Twitter	1
1.1.1. Trending topic	2
1.1.2. Twitter en Chile	3
1.2. AmorTv	3
2. Marco Teórico	5
2.1. Visiones referente al <i>gatekeeping</i>	7
2.2. Twitter y su relación con el periodismo	12
3. Estado del Arte	13
3.1. Estudios relacionados	13
3.1.1. Clasificación de tweets y usuarios de Twitter	13
3.1.1.1. Clasificación de usuarios	13
3.1.1.2. Ranking de enlaces compartidos en Twitter	16
3.1.1.3. Mecanismo de Ranking en Twitter como foro	19
3.1.1.4. Twitter para la recomendación de noticias	20
3.1.1.5. Clasificación de tweets orientada al usuario: Un enfoque de filtrado para microblogs	21

3.1.1.6. TURank: Clasificación del usuarios de Twitter basado en el análisis de un grafo usuario-tweet	23
3.1.2. Geolocalización de usuarios	25
3.2. Herramientas y plataformas relacionadas	28
3.2.1. Geofeedia	28
3.2.2. Paper.li	29
3.2.3. The Tweeted Times	30
3.2.4. FlipBoard	31
3.2.5. Summify	32
3.2.6. TwitterFall	34
3.2.7. Storyful	35
3.2.8. Wikipulse	39
4. Definición de la solución	44
4.1. Objetivos	45
4.1.1. Objetivo principal	45
4.1.2. Objetivos Secundarios	45
5. Propuesta	46
5.1. Arquitectura de la solución	46
5.2. Plataformas y herramientas utilizadas	47
5.3. Características del servidor	51
5.4. Modelo de Datos	53
5.4.1. Análisis de la línea editorial del medio objetivo	53
5.4.2. Geolocalización de usuarios	55
5.4.2.1. Implementación	55
5.4.2.2. Resultados obtenidos	55
5.4.3. Captación de usuarios	57
5.4.3.1. Captura de los medios de prensa (MP)	57
5.4.3.2. Captura followers de los medios de prensa (FMP)	58
5.4.4. Captura de tweets	61
5.4.5. Procesamiento de los tweets	62

5.4.5.1.	Definición del tópico	62
5.4.5.2.	Obtención del conjunto de tweets relacionados al tópico	63
5.4.5.3.	Depuración de conjunto de tweets relacionados al tópico	63
5.4.5.4.	Orden geográfico	64
5.4.5.5.	Orden de relevancia	65
5.4.5.6.	Panel de enlaces	66
5.4.5.7.	ON/OFF Medios de prensa	67
6. Evaluación y discusión		68
6.1.	Vistas del prototipo	68
6.2.	Caracterización de la población de datos capturados	72
6.3.	Resultados	77
6.3.1.	Referencias a noticias	78
6.3.1.1.	Medio de prensa Modelo	78
6.3.1.2.	Referencia a noticias	79
6.3.2.	Casos de prueba: El aborto	83
6.3.2.1.	Procesamiento del tópico	83
6.3.2.2.	Análisis de muestra representativa	84
6.3.2.3.	Análisis de contenido	84
6.3.2.4.	Análisis temporal	86
6.3.2.5.	Análisis de re-tweet	88
6.3.2.6.	Análisis geográfico	90
7. Conclusiones		91
7.1.	Conclusiones	91
7.2.	Trabajo futuro	94
8. Anexo		98
Bibliografía		99

Índice de figuras

2.1fig	Representación del proceso de gatekeeper [78]	6
3.1fig	Imagen del mapa interactivo de geofeedia donde en una zona delimitada por el usuario se reciben todos los feeds de los medios sociales.	29
3.2fig	Imagen del panel de noticias de geofeedia, donde cada columna recoge los feeds para ubicaciones geográficas distintas.	29
3.3fig	Vista principal del periodico Paper.li de un usuario de la plataforma	30
3.4fig	Vista de fotografías del periodico Paper.li de un usuario de la plataforma	30
3.5fig	Vista principal de un periodico creado en The Tweeted Times	31
3.6fig	Vista del formato de revista en Flipboard para visualizar los feeds de Twitter, donde se resaltan los tweet que han tenido mayor repercusión en la red y los enlaces compartidos	32
3.7fig	Vista de la recopilación de Summify	34
3.8fig	Vista de la recopilación de Twitterfall	35
3.9fig	Vista principal de Storyful	36
5.1fig	Diagrama conceptual de la arquitectura de la solución	47
5.2fig	Modelo de base de datos	53
5.3fig	Mapa con los usuarios por ubicación geográfica	56
5.4fig	Diagrama conceptual con las etapas para la captación de usuarios.	57
6.1fig	Vista nueva búsqueda tópico	68
6.2fig	Vista resultados tweets para nuevo tópico	69
6.3fig	Vista de tópicos	70

6.4fig	Vista de orden geográfico	70
6.5fig	Vista de orden por relevancia	71
6.6fig	Vista de orden temporal	71
6.7fig	Vista de links	72
6.8fig	Vista de tópicos	72
6.9fig	Distribución de tweets por hora de emisión	73
6.10fig	Distribución porcentual de tweets por hora de emisión según región	74
6.11fig	Tweet que generó nuevo record del mensaje más re-tweeteado.	75
6.12fig	Pneedimiento para determinar si un tweet se refiere a un hecho noticioso	80
6.13fig	Pneedimiento para verificar si el hecho noticioso haya generado una nota del medio modelo	81
6.14fig	Distribución de cantidad de tweets por día	86
6.15fig	Distribución de cantidad de tweets por periodos de quince días	87
6.16fig	Cantidad de re-tweets por quincena	88
6.17fig	Distribución geográfica de los usuarios	90

Índice de Algoritmos

1algo	Obtención de las palabras más frecuentes del timeline de un conjunto de tweets	54
2algo	Reconocimiento de ubicación del usuario mediante Levenshtein	55
3algo	Construcción lista de medios	58
4algo	Captura de usuarios	59
5algo	Gathering Followers	60
6algo	Gathering Followers con mejora	60
7algo	Algoritmo para la captura de tweets.	61
8algo	Obtención del conjunto de tweets relacionados al tópico	63
9algo	Clear: Clasificador	64
10algo	Origen Geográfico	64
11algo	Origen Relevancia	66
12algo	Obtención de enlaces externos contenidos en los tweets	66
13algo	Diseño de algoritmo para determinar si un tweet se refiere o no, a un hecho noticioso	81
14algo	Diseño-código para verificar si el hecho noticioso haya generado una nota del medio modelo.	82

Capítulo 1

Introducción

1.1. Twitter

Twitter es una red social que permite a los usuarios y usuarias enviar y leer mensajes cortos de un máximo de 140 caracteres, fue lanzada por una empresa de diez jóvenes en San Francisco en julio de 2006. Posee una interfaz web donde se muestran los propios mensajes del usuario(tweets). Los usuarios pueden indicar si desean que sus tweets sean públicos (que puedan ser vistos por cualquier usuario o usuaria de internet) o reservados a un ambiente privado (que sólo pueden ser visualizados por los seguidores y seguidoras del usuario o usuaria) y publicarlos mediante la plataforma web, mensajes de texto, la aplicación móvil o mediante clientes que facilitan el servicio.

Los usuarios y usuarias de Twitter siguen a otros y otras y son seguidos por otros y otras. A diferencia de la mayoría de los sitios de redes sociales en línea, como Facebook o MySpace, la relación de seguir y ser seguido no requiere reciprocidad.

Una práctica habitual en Twitter es responder a un tweet mediante un lenguaje simbólico de etiquetas bien definido: RT significa re-tweet (o replicar el mensaje en cuestión), "@"seguido de la dirección del identificador del usuario o usuaria se refiere a un usuario o usuaria en específico y un '#' seguido de una palabra representan un *hashtag* (Una forma de etiquetado de metadatos, los *hashtags* proporcionan un medio de agrupar este tipo de mensajes, ya que uno puede buscar el *hashtag* y obtener el conjunto de mensajes que lo contienen). El mecanismo de re-tweet permite a los usuarios y usuarias difundir la información de su elección más allá del

alcance de los seguidores del autor o autora original.

Twitter actualmente cuenta con cerca de 316 millones de usuarios y usuarias activas mensuales, donde se producen cerca de 500 millones de tweets al día [73] y presentando en el 2012 presentaba el impresionante ritmo de crecimiento de 11 nuevas cuentas de Twitter por segundo [47].

Twitter se clasifica como un servicio de microblogging, permitiendo a los usuarios y usuarias enviar actualizaciones de texto breves o micromedios como fotografías o clips de vídeo y posee una importante componente de tiempo real. Los usuarios y usuarias escriben mensajes de Twitter varias veces en un sólo día. Pueden saber lo que otros y otras están haciendo o pensando de manera inmediata, los usuarios y usuarias vuelven repetidamente al sitio y comprueban para ver lo que otros y otras están haciendo. De esta manera surgen numerosos informes de diversos eventos no sólo de la vida privada de cada usuario o usuaria, sino también de eventos públicos compartidos o vividos por muchos usuarios y usuarias.

En quizás el primer estudio del uso de Twitter, Java [45] identifica tres categorías principales de los usuarios y usuarias de Twitter: Las *fuentes de información*, los *amigos* y los *solicitantes de información*. Las *fuentes de información* publican noticias y tienden a tener una gran base de seguidores, estas fuentes pueden ser personas físicas o servicios automatizados. Los *amigos* es una categoría amplia que abarca a la mayoría de los usuarios y usuarias, incluyendo la familia, compañeros y compañeras de trabajo y extraños o extrañas. Por último, los *solicitantes de información* tienden a ser usuarios o usuarias que pueden crear información, pero que rara vez son seguidos por otros usuarios y usuarias de manera regular.

Java [45] también identifica varias categorías de tipos de usos de Twitter incluyendo la categoría de charla cotidiana, donde los usuarios y usuarias discuten acontecimientos de sus vidas personales o de sus pensamientos actuales, intercambian información o enlaces, noticias las que incluyen comentarios sobre la actualidad. El uso más relevante es la intención de conversación, considerando la aparición de la señal “como un indicador, Java determina que el 21 % de los usuarios y usuarias utilizaron Twitter para este propósito.

1.1.1. Trending topic

Twitter sigue y destaca las frases, palabras y *hashtags* que más a menudo son mencionados en la red social y los cataloga con el nombre de *trending topic* (tendencia o tema del momento).

Twitter muestra de manera predeterminada una lista de los diez *trending topics* del momento en una barra lateral situada a la derecha de la página principal de cada usuario y usuaria, a menos que se disponga lo contrario. Esta característica ha tenido gran repercusión en la prensa logrando que sea utilizado también para denominar un tema de gran interés. Algunos ejemplos de *trending topics* en Twitter fueron por ejemplo *la muerte de Michael Jackson*, *la muerte de Amy Winehouse*, *los finales de la UEFA Champion League* o *la aparición del nuevo Iphone*. Debido a las limitancias del largo de los textos en Twitter las temáticas son expresadas por frases principales como por ejemplo #QEPDMickaelJackson.

1.1.2. Twitter en Chile

Twitter cuenta con 5 millones de cuentas activas en Chile (posicionándose el 2012 entre los *top ten* en su uso a nivel mundial). Contando con un tiempo estimado de navegación y uso que fluctúa entre 7.7 horas diarias [62].

Según el perfil de uso de Twitter en Chile desarrollado en [46] la frecuencia de uso de Twitter corresponde a: 76 % la utilizan varias veces al día, 16 % al menos una vez al día y el resto al menos una vez por semana. Mientras que los horarios de mayor uso se encuentran desde las 10:00 hrs. en adelante (un 60 % de los usuarios y usuarias comienzan a utilizar la red a esta hora) alcanzando su *peak* de uso entre las 19:00 y 22:00 horas con un 73 % de los usuarios y usuarias.

Respecto a la intención de uso, 45 % de los usuarios y usuarias utilizan Twitter para mantenerse informado, 25 % para debatir y expresar opiniones, 16 % por entretenimiento y el restante 14 % se reparte entre diversas razones: mantener vínculos profesionales, mantener contacto con amigos y conocidos, para hacer nuevos amigos entre otras.

1.2. AmorTv

AmorTV [2] es un medio de prensa audiovisual estudiantil fundado a finales de las movilizaciones sociales del año 2011. Busca informar a la comunidad universitaria de la UTFSM sobre los principales hechos noticiosos del acontecer político y social.

Su línea editorial se basa principalmente en una mirada crítica anti-capitalista y pro-educación

gratuita (por su centralidad en lo universitario), con foco principal en conflictos y protestas sociales, tanto a nivel nacional como internacional.

AmorTV es el medio local de prensa que inspira la investigación y realización de este prototipo como nicho real y práctico, de la problemática de generar contenido informativo para una comunidad concreta de personas.

Capítulo 2

Marco Teórico

Una noticia es la comunicación de información seleccionada sobre un evento actual que es presentado posteriormente a través de cualquier medio de comunicación existente.[63]

El periodismo es un método de investigación y el estilo literario utilizado en la representación social y cultural de noticias. Sirve el propósito de jugar el papel de una maquinaria de servicio público en la difusión y análisis de las noticias y de la información. [38]

En el proceso de generación de una noticia, los y las periodistas son bombardeados con información provenientes de muy diversas fuentes de la cual, deben seleccionar y dar forma a la pequeña cantidad (de información) que se convierte finalmente en la noticia. Este proceso sería imposible sin las etapas de selección, redacción, edición, posicionamiento, programación, repetición y de cualquier otro tratamiento adicional de la información para convertirla en noticia.

Este conjunto de etapas se denomina como proceso de *gatekeeping* (y a quien los gestiona *gatekeeper*). Desde este proceso, se ofrece una imagen del mundo al público receptor de la noticia, por lo cual, es de vital importancia entender el proceso de *gatekeeping* y su impacto real en la realidad de los receptores y receptoras.

El *gatekeeping* es una de las teorías más antiguas proveniente desde las ciencias sociales, adaptada y desarrollada para su uso en el estudio de las noticias desde la década de 1950, enfocándose principalmente tanto en el producto producto final como en la información seleccionada o rechazada en cada etapa.

La figura 2.1 ilustra un conjunto de elementos que entran en el proceso de *gatekeeping*. No

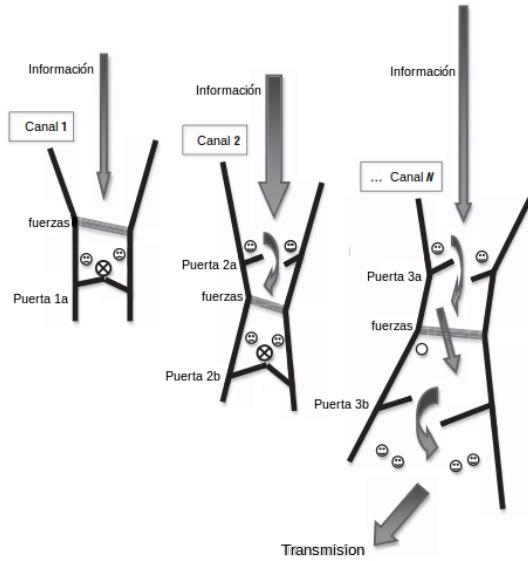


Figura 2.1: Representación del proceso de gatekeeper [78]

todos los elementos son seleccionados, algunos hacen su camino en los canales, que a veces se dividen en secciones, cada uno de los cuales se pueden introducir solamente pasando a través de una puerta. Diversas fuerzas facilitan o dificultan el flujo de artículos a través de las diversas puertas, mediante la variación en la magnitud y la dirección del canal. La Figura 2.1 muestra tres canales y muchos elementos de información, pero sólo un elemento se abre camino a través de un canal y se transmite a una o más audiencias. Las fuerzas negativas detienen el progreso de algunos elementos a través de los canales (es importante tener en cuenta que existen fuerzas externas, tanto antes como después de las puertas).

El elemento final presentado como el resultado del proceso de *gatekeeping* no sólo es el resultado de la selección, sino también el resultado de muchas otras fuerzas, ya que pasa a través de los diversos canales, secciones y puertas. Algunos procesos importantes están ocultos: Es el *gatekeeper* quien controla si la información pasa a través del canal y el resultado final. Cabe señalar que los *gatekeeper* toman muchas formas (ejemplo: personas, códigos profesionales de conducta, políticas de la empresa, los algoritmos informáticos, etc.) y todos toman decisiones, pero con distintos grados de autonomía. La autonomía varía desde idiosincrasias de una persona a conjuntos de reglas inquebrantables interpretados por algún algoritmo de un computador.

Las flechas varían en tamaño para indicar cómo los artículos cambian a medida que pasan a través de las diversas puertas o barreras. El elemento listo para ser transmitido es el resultado

de muchas influencias y no siempre se parece al artículo original.

2.1. Visiones referente al *gatekeeping*

Existen diversos análisis y críticas respecto al efecto y real impacto del proceso de *gatekeeping* en la generación de una noticia.

White[79] se refiere al proceso como muy subjetivo puesto que las diversas etapas dependen de juicios de valor basados en el conjunto propio de experiencias, actitudes y expectativas del *gatekeeper*. En investigaciones posteriores White sostiene que la selectividad de los y las periodistas es la principal fuente de “sesgo” de las noticias gatillado, entre otras cosas, por la necesidad de reducir una multitud de acontecimientos ocurridos en el mundo real a un número modesto, en un tiempo reducido.

Breed en su investigación sobre el control social en las salas de redacción [5] identifica a los y las editoras de los periódicos como los *gatekeeper* de facto que operan a través de medios indirectos para asegurar que sólo las noticias en consonancia con la política de la organización sean las generadas. Breed agrega “Las políticas de noticias podrían modificar o enterrar un evento y de esta manera denegar una información importante a la ciudadanía”.

Breed muestra cómo el *gatekeeper* más importante puede no ser necesariamente quien está relacionado más directamente con la selección ya que puede residir en otro lugar, dentro de los niveles más influyentes de la organización. Si la noticia es lo que el periodista dice que es, la subjetividad del *gatekeeper* problematiza profundamente el proceso de información. Reese y Ballinger en [59] argumentan que la razón radica en la expectativa de que actúe adecuadamente en nombre de la comunidad, el *gatekeeper* “ve a ella (a pesar de que nunca sea consciente de ello) que la comunidad oirá como un hecho único aquellos eventos que el periodista, como representante de su cultura, cree que es verdad”. Al igual que White, Breed señala que el proceso *gatekeeping* podría funcionar cumpliendo las expectativas de la comunidad a través de los códigos periodísticos y otras orientaciones, fuera de la influencia indebida de las y los editores.

Herbert Gans en [32] identifica las fuentes de poder dentro de la organización periodística y los incentivos que tienen las y los periodistas para cumplir con las normas del grupo y seguir las consideraciones prácticas. Gans señala que la construcción de la noticia no está principalmente

en el periodista o en el editor o editora, sino en el proceso por el cual todas las partes, las rutinas y las disposiciones de la organización se dedican a la creación de noticias. Reduciendo la responsabilidad directa respecto a la distorsión de individual de cada periodista.

Para Gans, el proceso de *gatekeeping* es el proceso de solución de los problemas relacionados con el envasado del flujo diario de los acontecimientos, en un producto comercial para el público. Para solucionarlo, los periodistas utilizan “consideraciones” para ayudar en el proceso, que debe ser aplicables sin demasiada deliberación, éstas deben ayudar a evitar la incertidumbre excesiva, ser flexibles, fácilmente racionalizadas o explicables a los demás y eficientes, garantizando los mejores resultados con el menor esfuerzo.

Lewin [49] plantea que las y los individuos deben entenderse en el contexto de cuatro sistemas: un microsistema (contexto inmediato), mesosistema (nexo de contextos inmediatos), exosistema (instituciones externas) y macrosistema (cultura o sistema social). Estos cuatro niveles aplicados a la redacción de una noticia incluyen el nivel individual de cada periodista, el nivel de las rutinas o prácticas de periodismo, el nivel de organización, el nivel de los medios de comunicación y el nivel del sistema social. Lewin sostiene además que en todos los niveles existen diversas fuerzas que determinan cuales elementos se convierten en noticias y cuales no, limitando la autonomía de los *gatekeeper* y dando forma a las noticias de manera consistente.

Lippman en [50] señala taxativamente que el papel de la prensa es el de ser en cierto modo servidor y guardián de las instituciones y sugiere que las fuerzas planteadas por Lewin se relacionan con este rol que juegan los procesos de comunicación y los medios. Según Laswell en [48] este proceso realiza tres funciones:

1. Vigilancia del entorno, revelando amenazas y oportunidades que afecten a la posición de valor de la comunidad y de las partes que la componen.
2. Correlación de los componentes de la sociedad en cuanto a dar una respuesta al entorno.
3. Transmisión del legado social.

De una forma esquemática ha sido descrito este triple papel de esta otra manera: vigilancia, foro para la discusión y escuela.[1]

Smith [43] explica precisamente una de estas tres funciones capitales, la de vigilancia del

entono mediante la metáfora del *perro guardián*¹ señalando además que las y los periodistas, en cuanto críticos y críticas sociales y políticas, no desempeñan correctamente la labor encomendada a causa de carencias estructurales en estos cuatro aspectos:

1. El ejercicio periodístico es básicamente una actividad de escaso rigor intelectual y con marcada tendencia a la simplificación.
2. Las y los periodistas suelen carecer de conocimientos técnicos adecuados para la mayor parte de las cuestiones complejas de la vida actual.
3. El trabajo periodístico se ejecuta sin la reflexión y el sosiego que son deseables en una adecuada labor crítica.
4. Es evidente la falta de una actitud juiciosa y equilibrada en la mayor parte de las y los periodistas, que renuncian a hacer un balance de los datos positivos y negativos para reducirse únicamente a una esquemática y simplificadora enumeración de defectos aparentes sin analizar las causas.

Jean-François en [60] critica la tendencia a priori que inducen los distintos medios de prensa en las distintas noticias.². Real, Agudiez, Príncipe en [81] afirman que actualmente existe un descontento y desilusión ciudadana con los medios de prensa pues éstos últimos no cumplieron con su parte del contrato social (la de velar por la transparencia y la difusión de ésta información).

Con una visión contraria Lorenzo Gomis en [34] desarrolla la siguiente idea: los medios de comunicación y los periodistas no sienten interés por los problemas derivados de las posibles repercusiones de sus mensajes, pues no son ellos quienes los generan intencionalmente, sólo

¹ "...la tarea del *perro guardián* suponía que cada periódico atacaría y defendería desde una posición determinada, y que del conflicto entre todos, surgiría la verdad (...) Hoy en día, por el contrario, la prensa abarca muy numerosos medios de noticias, tanto impresos como electrónicos. En esta categoría se incluye un escogido y selecto puñado de diarios y semanarios, los servicios cablegráficos, las noticias de redes radiofónicos y, sobretodo, las principales cadenas de televisión. En virtud de sus dimensiones y prominencia, son estos pocos medios de élite los que establecen el tono de los programas de cobertura para la mayor parte de la prensa"

²"...la generación de noticias debe resultar de la información correctamente establecida, y no dirigir la elección de esa información a impulsos de un prejuicio selectivo, que metamorfosea la despiadada ferocidad para con unos en indulgencia y sin límites con otros (...) lo que predomina desgraciadamente en muchos periódicos de nuestro entorno sociocultural es el dirigismo apriorista en contra del poder, la predisposición condenatoria contra los actos emanados de las instituciones gubernamentales "

los comunican y los efectos de ambas acciones son incomparables³. Para este autor el accionar de los *gatekeeper* no tiene motivaciones otras que no sean técnicas: “Los seleccionadores o gatekeepers no ponderan la influencia potencial de los hechos en cuanto a sus efectos políticos o sociales, sino que consideran únicamente su condición técnica de noticia y, en caso de duda, es sumamente probable que la noticia quedará sin publicar”.

A finales de la década de los noventa surgió una alternativa a esta visión de los medios de prensa como una entidad que resguardaba de forma neutral los intereses de las y los ciudadanos, que implica a los mismos ciudadanos y ciudadanas, esto posibilitado -entre otras cosas- por la irrupción de Internet y las profundas transformaciones que supone en la información periodística y las vías de acceso a la información.

Las nuevas tecnologías brindan diferentes oportunidades para la incorporación de las inquietudes de las y los ciudadanos en los discursos dominantes de los medios mediante su participación directa en la producción informativa. Un nuevo escenario donde los pasivos y silenciosos ciudadanos se convierten en potenciales productores de información.

La adaptación de los medios de comunicación a este nuevo escenario, prevé una comunicación bidireccional entre medio y audiencia. En ese sentido, los medios de comunicación intentan consolidar mecanismos de participación en sus medios digitales. No obstante, tal y como señala Hermida en [39] en una investigación que examina las oportunidades que la audiencia tiene de participar en el proceso periodístico (participación, acceso, selección, edición y distribución) identifica que los medios de comunicación suelen ofrecer herramientas de participación similares entre sí y raras veces permiten que participe del *gatekeeping*⁴.

Shayne Bowman y Chris Willis presentan en su informe “*We Media*”[4] las valiosas ventajas de incorporar a las y los ciudadanos en la producción de información:

1. La posibilidad de pedir sugerencias y correcciones al público.
2. La posibilidad para las lectoras y lectores de que hagan comentarios.

³ Gomis lo explica de la siguiente manera: ”...La mayor influencia que se ejerce en los medios no es a través de los comentarios, sino de los mismos hechos. Y por lo tanto influye quien aporta el hecho, ya sea el interesado en el hecho que le favorece, o ya sea el interesado que perjudica a su adversario. Los medios son en definitiva la escena donde se luchan los productores de hechos para influir en la gente, mientras que los que controlan el medio sólo se interesan relativamente en esa pugna (...) Los más interesados en influir en los medios no son ni los que los poseen ni los que trabajan en ellos. Curiosa situación”

⁴”Nuestro estudio indica que la selección, o filtro, es el proceso de *gatekeeping* más cerrado para los usuarios, y creemos que seguirá siéndolo” [39]

3. La función de un filtro de noticias para noticias encontradas en la web a través de enlaces.
4. El control de exactitud en la información publicada.
5. El enriquecimiento de fuentes e ideas para periodistas gracias a las sugerencias e historias presentadas por las y los lectores.

Existe una gran cantidad de situaciones en las que las y los ciudadanos han colaborado con el envío de informaciones, imágenes y vídeos sobre un acontecimiento. Contribuyendo a comunicar un suceso desde su propia perspectiva. A ese fenómeno se le reconoce como periodismo ciudadano, periodismo de fuente abierta o periodismo en red, que son sinónimos de lo que Bowman y Willis denominan periodismo participativo en Internet. En [4] lo definen como “el acto de la ciudadanía que juegan un papel activo en el proceso de colectar, reportar, analizar y diseminar información”, que con frecuencia ocurre en un medio social y colaborativo.

En este proceso hay una comunidad de internautas que se reúnen para producir de forma colaborativa. En este caso, no existen límites geográficos, lo que cuenta es el conocimiento, el trabajo creativo y el deseo de colaborar. Esa comunidad de usuarios también se dedica a filtrar contenidos y recomendarlos a sus contactos en la Red. A este proceso Bruns en [6] denomina *gatewatching*. Que a diferencia del proceso de *gatekeeping* no se busca seleccionar la información, sino de dar facilidades y atajos de lectura.

Los formatos más comunes para la participación en los medios de comunicación incluyen: Blogs de ciudadanos y ciudadanas, envío de fotografías y vídeos, envío de textos, entrevistas colectivas, comentarios, ranking de contenidos elaborados de acuerdo con los votos de usuarios, foros, blogs de periodistas, encuestas y comentarios en redes sociales como Facebook y Twitter [39].

Arriagada y Navia en [80] plantean que la irrupción de la tecnología no solo amplia las posibilidades de producción y acceso de la ciudadanía sino también modifica las relaciones de poder e influencia entre los medios, las audiencias y los grupos políticos, empoderando a las y los ciudadanos permitiendo el desarrollo de interacciones entre la clase política y la gente de forma independiente del quehacer de los medios⁵.

⁵ Arriagada y Navia lo explican de la siguiente manera: "Si históricamente los medios podían ser considerados como el cuarto poder, en tanto constituyan una institución de la sociedad que podía vigilar el comportamiento de la clase gobernante, ahora los propios medios son a su vez vigilados por grupos que, a través de redes sociales online, participan activamente en los debates públicos, y que a menudo presentan niveles de desconfianza relativamente

2.2. Twitter y su relación con el periodismo

Además de red social, Twitter se ha convertido en un sistema de noticias compartidas en línea, que se basa en mensajes cortos, rápidos y frecuentes. Hermida [40] describe este sistema como un medio ambiente que muestra la información en un espacio ocupado por la o el usuario. En este sistema, las y los usuarios reciben la información en la periferia de su conciencia, es decir, Twitter no requiere la misma atención cognitiva que un e-mail, por ejemplo. Se refiere a una especie de medio ambiente periodístico que ofrece diversos medios para recopilar, comunicar y compartir noticias e información.

Debido al acceso a Twitter desde teléfonos móviles existe una nueva estructura de movilidad al periodismo, tanto en la producción como la difusión de la información, ya que la conectividad de los dispositivos móviles permite la instauración de la instantaneidad de la noticia. Siendo así, la movilidad reconfigura el trabajo de edición de blogs y la generación de contenidos periodísticos en las redes sociales.

Por su carácter instantáneo, Twitter sirve como plataforma para la realización de la cobertura periodística de eventos, en la cual se elabora un tipo de crónica de última hora o flash⁶, pero en orden cronológico inverso.

Un guía oficial de Twitter para redacciones periodísticas fue lanzado en junio de 2011, en ella se señala: "Twitter es una herramienta de todos los periodistas pueden utilizar para encontrar las fuentes más rápido, contar historias mejor y construir una mayor audiencia para su trabajo"[75].

altos hacia las instituciones. En las redes sociales online, estos niveles de desconfianza pueden alcanzar también a los medios. Es más: en la medida que la gente percibe a los medios como parte de la institucionalidad o como representantes de las élites, los crecientes niveles de desconfianza en las instituciones también pueden alcanzarlos."

⁶El flash corresponde a una información de última hora, de elevada importancia y gran impacto informativo. Suele ser un texto conciso en que se priorizan los aspectos más relevantes del acontecimiento. De acuerdo con Salaverría el flash es solo un arranque de una cadena de informaciones, que resultará en un texto más completo que responda a las seis preguntas clásicas de toda noticia. El flash se ha ido convirtiendo en notas informativas cortas, limitadas a 140 caracteres (límite establecido por los servicios de mensajería instantánea y Twitter).[61]

Capítulo 3

Estado del Arte

3.1. Estudios relacionados

3.1.1. Clasificación de tweets y usuarios de Twitter

Debido a la gran cantidad de usuarios que forman hoy en día la red social de Twitter, la clasificación tanto de usuarios como de los tweets se ha convertido en una interesante materia de estudio para la comunidad científica. En el siguiente apartado se revisan los trabajos realizados referentes a ésta temática.

3.1.1.1. Clasificación de usuarios

En el trabajo realizado por Marco Pennacchiotti y Ana-Maria Popescu en [54] se realizó una clasificación del perfil de los usuarios considerando principalmente cuatro aspectos diferentes:

1. *Perfil del usuario:* En la actual API Twitter v1.1 permite el acceso a la información básica del usuario como el nombre de usuario, la locación, una pequeña biografía además del número de seguidores, el número de personas a las que sigue el usuario y el número de tweets. En trabajo anteriores Cheng en [8] estableció que sólo el 26 % de los usuarios de Twitter reportan su ubicación como una ciudad específica, el resto provee locaciones generales (estados o países) o lugares imaginarios. Por su parte Marco Pennacchiotti y Ana-Maria Popescu tras analizar un corpus de 14M de cuentas de usuarios, encontraron que sólo el 48 % provee una biografía corta y 80 % una ubicación, de cuya información se

intentaron determinar el género del usuario y su etnicidad pero los resultados fueron de muy baja calidad. Mediante una muestra 15000 usuarios elegidos de forma aleatoria y un conjunto de editores, se les pidió identificar la etnicidad y el género a partir de la imagen del avatar de twitter obteniendo que menos del 50 % de las imágenes se correlacionaba de manera clara con alguna etnia mientras que el 57 % se correspondía con algún género. También encontraron que las imágenes podían ser engañosas: en el 20 % de los casos la imagen no era del dueño de la cuenta sino de una celebridad o de otra persona.

Con toda esta información estadística concluyen que los campos del perfil del usuario no contienen suficiente información de buena calidad para ser utilizada para una clasificación.

2. *Comportamiento para escribir tweets:* Esta determinado por un conjunto de métricas de las interacciones existente entre la red social y el usuario: el número promedio de tweets por minuto, número de respuestas, entre otras. Han existido trabajos previos sobre estos aspectos. En [45] han ratificado percepciones intuitivas como: Cuando los usuarios posteán rara vez pero tienen una gran cantidad de seguidores tienden a ser un buscador de información. Mientras que los usuarios que poseen URL's en sus propios tweets son proveedores de información. Rao en [58] sugiere que el estudio de estas componentes no es útil para la mayoría de las tareas de clasificación y que queda incluida en los rasgos lingüísticos.
3. *Contenido lingüístico de los tweets:* el contenido lingüístico de los tweets encapsula los temas principales de interés del usuario, así como su uso de léxico. La información lingüística es útil para realizar una clasificación de los usuarios según su estilo de escritura tales como textos oficiales, blogs, conversaciones o traducciones.

El análisis se hace mediante el uso de *palabras prototípicas* las cuales son expresiones típicas de las personas pertenecientes a una clase específica, así como frases relacionadas a intereses típicos de dicha clase.

En este estudio se realiza un análisis de los hashtag prototípicos, basados en la hipótesis de que si los usuarios de una clase están interesados en los mismos temas, los temas más populares de esta clase se pueden encontrar mediante la recopilación de estadísticas sobre

hashtags usados. A modo de síntesis: Se representa a un usuario mediante el conjunto de palabras de sus tweets y mediante éstas se intenta clasificar a dicho usuario.

También se realiza un análisis de sentimiento a los tweets, pues frente a un tema en particular dos clases pueden expresar distintos sentimientos. Por ejemplo dos grupos políticos pueden expresar opiniones positivas o negativas de una figura pública dependiendo generalmente si es del mismo sector político a ellos. Concretamente se realiza recogiendo un conjunto de palabras para las clases observadas en este estudio sobre las que el usuario particular tiene una opinión global que en su mayoría no es compartida por otra clase diferente.

4. *Información de Twitter*: Estas características exploran las relaciones sociales establecidas por el usuario con los demás que él sigue, a quien le responde o que personas re-tweetea. Existe la idea intuitiva de que las personas pertenecientes a una clase son más propensos a seguir las cuentas de ciertas personas y a responder a ellas (por ejemplo, las jóvenes pueden tender a responder a la cuenta de Justin Bieber)

Las clases específicas que se clasificaron en este trabajo son: afiliación política, clasificación de si es o no seguidor de Starbucks y si los usuarios pertenecen o no a la etnia afroamericana. Los resultados obtenidos son que para los dos primeros es posible conseguir un buen resultado con una considerable precisión, mientras que para el tercero es más compleja la tarea.

Para determinar la afiliación política se concluye que las mejores características para su determinación son lingüísticas y de perfil. Siendo desconsiderable el aporte que entregan a esta determinación la información de Twitter y el comportamiento en "tweetear".

Para determinar si una usuario pertenece a la clase seguidor potencial de Starbucks se puede hacer con precisión razonable (0.763) considerando la información del perfil y análisis lingüístico como las características más útiles para este objetivo.

Se concluye además que la relación entre los seguidores y amigos es también una característica relevante por si sola para determinar la clase de seguidores, pues sugiere que los aficionados de starbucks son los que siguen a los demás más que seguirse entre sí pues en su mayoría son los solicitantes de información, probablemente gente en busca de ofertas y cupones.

Si bien el uso del léxico determina en este experimento con bastante precisión a la etnia afro-americana por sus variadas expresiones, cabe destacar que dichas expresiones ligüísticas

han sido ampliamente adoptadas por otros grupos. Por lo cual, esta característica posee una clara limitante en su aplicación. Los mejores resultados para determinar la etnia de un usuario es mediante la combinación del léxico y si siguen a celebridades a fines. Se encontró también que la tarea de clasificación puede ser ayudada por información del perfil.

3.1.1.2. Ranking de enlaces compartidos en Twitter

En [21] Dong utiliza como fuente de información fresca los contenidos y enlaces compartidos en Twitter para las búsquedas web en tiempo real. Dong identifica que en vista de la investigación de Hughes en [42] donde concluye que ante un evento inesperado, los tweets contienen más información relevante que en una situación normal (y tienen un enfoque más de broadcasting), es posible considerar Twitter como una buena fuente de información en tiempo real en base a cuatro oportunidades:

- Los enlaces compartidos pueden corresponder a noticias o no. (permitiendo recoger información sobre los vínculos que no son noticias y mejorar los resultados de la búsqueda)
- Los enlaces difundidos son publicados en base a las distintas prioridades personales de los usuarios lo que aporta un interesante grado de diversidad.
- La red de Twitter permite realizar mediciones de autoridad a los creadores de tweets.
- Los tweets cuentan con metadatos relacionados que permiten clasificarlos e inferir en base a su relevancia.

Existen aspectos a considerar en el procesamiento de enlaces como por ejemplo: que los enlaces propagados en Twitter puedan corresponder a *spam* referente a un producto o marca sin contenido relevante, para limpiar los resultados de estos contenidos se utilizan heurísticas a fines a estos propósitos (como filtrar enlaces twitteadas por el mismo usuario más de dos veces o solo twitteadas por un mismo usuario). Se consideraron las siguientes características referentes a los enlaces y los tweets que los contenían:

- **Características textuales:** Estas características fueron analizadas por considerarse que las palabras que acompañan a una URL en un tweet pueden entregar información relevante sobre esta. Principalmente se realiza un conteo de estas palabras y la cantidad

de repeticiones existentes para todos los tweets analizados, generándose un conjunto de pares de palabras y url's relevantes. (Similar al análisis del *Contenido lingüístico de los tweets* considerado en el estudio analizado anteriormente [54]).

- **Características de redes sociales:** Se aplica el concepto de autoridad a los usuarios de Twitter, vinculados mediante las relaciones de re-tweet entre ellos.
- **Otras características:** Se define un conjunto de diez características adicionales, muchas de las cuales consideran el ranking establecido en base a la autoridad de los usuarios. Estas se dividen en tres grupos:
 - Referentes al promedio del conjunto de usuarios que publicaron la URL: número promedio de *followers* de los usuarios, número promedio de tweets de los usuarios, numero promedio de usuarios que retweetean los tweets que contienen la URL, promedio de usuarios que responden los tweets que contienen la URL, promedio del número de personas a las que siguen los usuarios, promedio del ranking de autoridad definido anteriormente.
 - Referentes al usuario que inicialmente twitteó la URL: Se asume autoridad para el primer emisor de la URL en base a algunos criterios como número de *followers*, número de tweets, número de usuarios que realizaron retweet, número de respuestas, número de personas a las que sigue.
 - Referentes al usuario que posee mayor ranking de autoridad Las características observadas son número de *followers*, número de tweets, número de usuarios que realizaron retweet, número de respuestas, número de personas a las que sigue y el ranking de autoridad.

El ranking como tal, fue elaborado en base a una máquina de aprendizaje de ranking (MLR) en la cual se consideran las características anteriores con distintas ponderaciones de importancia calificándolas con un factor en una escala de [0,100](de forma descendente, las más importantes son: La característica de repetición de la URL en los distintos tweeets y su relación con las palabras relacionadas, número de seguidores del usuario con mayor grado de autoridad, número de usuarios que retweetean la URL al usuario con mayor grado de autoridad,

número promedio de usuarios que retweetean los tweets que contienen la URL y número promedio de usuarios que siguen a quienes han twitteado la URL). Los resultados se clasificaron en base a una tupla ($query, URL, t_{query}$) con grados de relevancia, los resultados obtenidos fueron contrastados con la opinión de cinco expertos en cinco grados de clasificación de los pares URL-query: perfecto, excelente, bueno, justo y malo. Se agregó un sistema de etiquetas de tiempo debido al interés especial de esta dimensión del problema, donde los tweets fueron clasificados en dos grandes categorías: sin sensibilidad de tiempo y con sensibilidad de tiempo con las subcategorías: *reciente*, *de alguna manera reciente*, *de alguna manera fuera de fecha* y *totalmente fuera de fecha*.

Si los documentos son *sin sensibilidad de tiempo*, *reciente* o *de alguna manera reciente* al momento de su clasificación preservan su clasificación, en cambio, cuando son *sensibles al tiempo* se utiliza un sistema de descenso de categoría:

- *Descenso de un grado*: Si el resultado es *de alguna manera fuera de fecha* se realiza un descuento de un grado (por ejemplo de excelente a bueno)
- *Descenso de dos grados*: Si el resultado es *totalmente fuera de fecha* se realiza un descuento de dos grados (por ejemplo de excelente a malo)

Los resultados obtenidos fueron contrastados con expertos bajo el criterio de que sólo evaluarán una query como con contenidos relevantes de última hora realizada al sistema si el sistema arrojaba al menos un documento creado en las últimas 24 horas con contenido relevante. Se obtuvo que el 91,7 % de las consultas fueron clasificados de esta manera. Entre los resultados se observa que gran cantidad de URL's poseen la calidad de enlaces frescos, lo que hace concluir que casi no existen documentos obsoletos en las URL's de Twitter.

Del mismo caso, ocurre que el porcentaje de edades de los enlaces clasificadas como *perfectas* o *excelentes* es más alta comparativamente que los enlaces *regulares*, mientras que las clasificadas como *justo* y *mala* son más bajas que las URLs *regulares*, demostrando de esta manera que las características de Twitter pueden mejorar el rendimiento de un sistema de *ranking* sensible al tiempo.

Finalmente se concluye que Los enlaces propagados en Twitter son útiles para mejorar potencialmente la clasificación de consultas de búsqueda sensibles al tiempo.

3.1.1.3. Mecanismo de Ranking en Twitter como foro

En [13] se presenta un análisis de un sistema de ranking que podría ser perfectamente aplicado a Twitter por sus características, sugiriendo una arquitectura genérica de un sistema de clasificación para diversos items (tweets, post u otro) con la intención de conseguir la mejor clasificación posible con el menor esfuerzo implicado.

Las características deseadas para el sistema de ranking propuesto en este trabajo son:

- *Precisión del Ranking*: El ranking debe ser preciso aún cuando no sea posible re-evaluar nuevamente todos los items implicados.
- *Revisión del ancho de banda*: El ranking debe converger al orden correcto, dentro del nivel de precisión deseado, rápidamente con una pequeña cantidad de retroalimentación por artículo.
- *Baja Latencia*: Los usuarios no deben esperar mucho tiempo para recibir un estimado de sus puntuaciones o clasificaciones actualizadas.
- *Equidad*: Los items deben ser tratados igualmente con respecto a la clasificación y la revisión.

Debido a que la distribución de probabilidad con la cual son evaluados los diversos elementos depende principalmente del orden en que son presentados para su evaluación, se intenta contrarrestar este efecto mediante el diseño de formas explícitas de evaluación en este trabajo. Bajo esta misma perspectiva, se elige el método de evaluación comparativo entre dos elementos a modo de torneo que funciona principalmente de la siguiente manera: cuando el usuario envía un nuevo elemento (tweet, comentario, post) se le muestra un par de otros items (seleccionados dependiendo la distribución de torneo sorteada al azar) entre los cuales selecciona el mejor, dichas clasificaciones son recogidas y evaluadas, estas evaluaciones cuentan con una mejor estimación de rango entre más evaluaciones poseen dichos elementos.

Finalmente se concluye que el sistema de *ranking* planteado posee mejor rendimiento - evaluado en base a las características deseadas- que un sistema de calificación individual (como el sistema de clasificación por estrellas de plataformas como Netflix).

3.1.1.4. Twitter para la recomendación de noticias

Muchos de los sistemas actuales de recomendación, se basan en las preferencias personales de los usuarios, en [55] utiliza Twitter y fuentes RSS para este cometido, debido a su rapidez de actualización y gran volumen de actividad que posee la gran red de microblogging.

El sistema se compone de tres grandes componentes:

- Componente Web: Encargada de reunir la información disponible del usuario en RSS y en Twitter para recoger las preferencias del usuario.
- Index Lucene¹: Responsable de la indexación y la minería de la información obtenida.
- Recomendación: El motor genera una lista clasificada de historias RSS basado en la co-ocurrencia de términos populares dentro de los post y gustos del usuario en Twitter.

La indexación se genera principalmente al transformar las palabras contenidas en los últimos tweets del usuario en una matriz de co-ocurrencia M (M_{ij} representa la cantidad de ocurrencias que posee la palabra j en el tweet i), las co-ocurrencias mayores se relacionan con el conjuntos de artículos que las contienen, tras sumar esta cantidad se le asigna un puntaje basado en la sumatoria de co-ocurrencias evaluadas puntaje asignado. De esta forma se crea un ranking de temas y artículos a sugerir (ordenados entre estos mismos según grado de relevancia). El usuario puede elegir tres estrategias distintas de recomendación:

- Ranking Público: Basado en el análisis de los tweets públicos del timeline del usuario en Twitter.
- Ranking de los amigos: Basado en el análisis de los tweets públicos de los timelines de los amigos del usuario en Twitter.
- Ranking de Contenido: No utiliza Twitter, sólo considera las 100 mayores ocurrencias de palabras del análisis de las RSS.

Se mide la aceptación de las recomendaciones realizadas por este sistema mediante su evaluación con un pequeño grupo de 10 participantes, en un período de 5 días y se cuantifica la

¹Lucene es una biblioteca de búsqueda de texto completo extremadamente rica y poderosa escrita en Java

cantidad de clicks recibidos por noticia recomendada por el sistema. Cada participante configuró el sistema con su información correspondiente.

Al analizar los resultados se obtuvo una clara diferencia en el comportamiento de los usuarios cuando se comparan las estrategias basadas en Twitter a la basada en el contenido prede terminado. Se observa, por ejemplo, que para la primera prueba se recibe una media 8,3 y 10,4 *clicks* por cada usuario en comparación con sólo el 5,8 *clicks* por usuario en la estrategia basada en el contenido; expresando un relativo aumento de entre el 30 % y el 45 % para las estrategia basada en Twitter.

Se observa también que el *ranking* de mayor uso de las recomendaciones basadas en el *ranking de los amigos* en comparación a las del *ranking público*, resultado que se contradice con un cuestionario posterior realizado, donde un 67 % de los usuarios indicó su preferencia por *ranking público* mientras que sólo el 22 % de los usuarios señaló preferencia por *Ranking de los amigos*. Ninguno de los participantes se mostró partidario de l estrategia de *Ranking de contenido* y un 11 % no saben cual estrategia prefieren.

3.1.1.5. Clasificación de tweets orientada al usuario: Un enfoque de filtrado para micro-blogs

En [77] se plantea la relevancia de un filtrado de contenido en Twitter, ya que debido a su enorme volumen de información carece de valor sin no es presentada de manera adecuada.

Ibrahim y Bruce plantean que la acción de re-tweetear para clasificar a los usuarios posee un gran valor: El re-tweet incluye leer el tweet, decidir que vale la pena compartir y luego actuar sobre ella, por lo cual es posible considerar el re-tweet como una señal explícita de que el usuario considera el tweet como información relevante.

Basado en esta apreciación, se busca en este trabajo responder a la siguiente interrogante ¿Podemos clasificar a los usuarios basándonos si ellos re-tweetean un tweet específico?

El objetivo de este trabajo es clasificar la cuenta de un usuario para mostrar los tweets entrantes en un orden descendente en función de su probabilidad de ser retuiteado por el mismo. Una clasificación efectiva ayudará al usuario a encontrar los tweets potencialmente más interesantes.

Como experimento preliminar, se utiliza un árbol de decisión (J48) (implementado en WEKA), para ver con qué precisión es posible clasificar los tweets como re-tweetable o no para un

usuario específico. Para entrenar el clasificador, fueron utilizados cuatro grupos de rasgos:

- Basado en el autor: Se refiere a características que pueden ser deducidas a partir del perfil del usuario: ¿Es el usuario un autor de élite? (dividido en dos clasificaciones: élite local si su cantidad de seguidores esta en el rango de 10K-50K followers, élite global si su cantidad de followers es mayor a 50K), cantidad de followers, cantidad de personas a las que sigue, cantidad de tweets, edad del perfil (cantidad de días desde la creación de la cuenta), tasa de tweets, cantidad de favoritos, ¿tiene descripción?, ¿su idioma es el inglés?. Principalmente características relacionadas a que tan activo es el usuario y la autoridad que posee.
- Basado en los tweets: Se refiere a características sintácticas del tweet: La puntuación TF-IDF²), ¿contiene hashtag? ¿contiene urls? ¿Menciona a otros usuarios? ¿utiliza comillas para citar? ¿escribe el mismo carácter tres veces seguidas (por ejemplo: hoolaaa)? ¿utiliza emoticonos?, la cantidad de re-tweet del tweet, el largo promedio de los tweets. Algunas de estas características dan implicaciones sobre que tan bien está escrito el tweet, la categoría, la audiencia y la popularidad del tweet.
- Basado en el contenido: Se refiere a las características relativas al contenido del tweet: novedad del tweet (distancia coseno de términos de los otros tweets que aparecen en el timeline en la última semana), lo inesperado del tweet (distancia mínima de la distancia coseno de términos con los demás tweets del autor).
- Basado en el usuario: Características relacionadas a la cuenta del usuario que esta siendo clasificado en Twitter: Del tweet re-tuiteado ¿sigue el usuario al autor? ¿se utilizó el hashtag del tópico relacionado en otros tweets? ¿se compartieron enlaces con el tweet? ¿se menciona al usuario en el tweet en cuestión?

Para el conjunto de entrenamiento se realizaron dos clases: los tweets re-tuiteados y los que no son re-tuiteados. Se consideró además la clasificación de *usuario ordinario* si el usuario tiene de 10 a 1000 seguidores o personas a las que sigue, escribe entre 1 a 200 tweets por semana y ha twitteado más de 10 veces.

²TF-IDF es una estadística numérica que se pretende reflejar la importancia de una palabra en un documento, colección o corpus

Los resultados obtenidos señalan que ningún grupo de características arrojó resultados satisfactorios por si solos, la mejor características de todas es la tasa de tweets(cantidad de tweets por semana).

De las características basadas en los tweets las más valiosas fueron: si el tweet fue retuiteado y la puntuación tf-idf. De las características basadas en el usuario se considera útiles las características: ”¿es el autor del tweet una persona a la que el usuario sigue?”, ”¿se compartieron enlaces con el usuario?”, ”¿se utilizó el hashtag del tópico relacionado en otros tweets?”, ”¿estoy mencionando en el tweet?”.

3.1.1.6. TURank: Clasificación del usuarios de Twitter basado en el análisis de un grafo usuario-tweet

En [82] se aborda el problema de identificar los usuarios con autoridad en una red de microblogging como es Twitter. Se entiende por un usuario con autoridad aquellos usuarios que frecuentemente suben información a la red, que es considerada relevante y es propagada de forma rápida y amplia.

Debido a la gran cantidad de información que se encuentra en esta red social es preciso identificar los usuarios con autoridad que dinamizan y aportan con información relevante, su identificación podría tener múltiples impactos en el manejo y categorización de la información. Muchos trabajos enfocan su análisis en la estructura de vínculos de seguidores que mantienen los usuarios, sin embargo, la mayoría de los usuarios siguen de vuelta al usuario que comenzó a seguirlos por un acto de cortesía formal, por lo cual es poco relevante.

En este trabajo se plantea un algoritmo que considera la puntuación de autoridad de los usuarios de Twitter considerando un grafo de relaciones sociales además del análisis del flujo de tweets entre los usuarios (mediante el re-tweeteo).

En cuanto a la propagación de la información, cabe señalar que los re-tweets tienen distintas características dependiendo del objetivo que esta acción tenga, si es un re-tweet conversacional tiene sólo re-tweets de los implicados en dicha conversación, muy por el contrario si es un re-tweet de difusión, este tienen gran número de re-tweet y se propagan ampliamente. Por lo anterior, no es suficiente sólo contabilizar la cantidad de retweets para medir la autoridad de un usuario, ya que los retweets conversacionales son menos relevantes para la puntuación de autoridad en cuestión.

Se consideran además los re-tweet realizados por usuarios con autoridad, pues es probable que estos retweets son más útiles que retweets de usuarios que no son autoridad.

Se utiliza un tipo de grafo direccional llamado *esquema de transferencia de autoridad* donde se representan el dominio del discurso y los flujos respectivos de autoridad donde los nodos representan todo el conjunto de elementos de destinos y las aristas todo el conjunto de relaciones o trasferencias que puede ocurrir entre ellos. Este gráfico es evaluado mediante *ObjectRank*, que es una extensión de PageRank[52] para medir la importancia de los objetos en una base de datos teniendo en cuenta el tipo de aristas del grafo así como el tipo de nodo.

La evaluación de los distintos usuarios se basa en tres observaciones:

- Un usuario seguido por muchas autoridades probablemente sea también una autoridad
- Un tweet re-tweeteado por alguna autoridad probablemente sea un tweet útil
- Un usuario que posee muchos tweets útiles es probablemente una autoridad

Basado en estas observaciones fue construido un grafo donde los nodos representan usuarios y tweets mientras que las aristas representan las relaciones entre usuarios y tweets. Este grafo *usuario-tweet* permite comprender cómo se propaga la información entre los usuarios mediante el retweet. Posteriormente se realiza un análisis de enlaces y se calcula la autoridad de los usuarios utilizando *ObjectRank*.

Las pruebas realizadas fueron contrastadas por un conjunto de expertos, y se obtuvo como conclusión que el número de followers y la cantidad de re-tweet no son suficientes para determinar si un usuario posee o no autoridad. El cuantificador de re-tweet tiende a extraer a los usuarios que utilizan el re-tweet para las conversaciones con el fin de especificar al usuario al cual se dirigen, en estos casos los tweets no transmiten información útil.

En Twitter, un usuario sigue a otro, si es probable que el usuario transmita información útil incluso sin garantías. En muchos casos, los seguidores no se dejan de seguirlo incluso si resulta que no transmite más información útil. Esto ocurre porque los usuarios no recuerdan a todos los usuarios a los que están siguiendo y debido al gran número de cuentas a las que el usuario sigue (de 100 a más de 1000 en muchos casos). Incluso si un usuario tiene una gran cantidad de seguidores no implica que estos sean frecuentemente re-tweeteados.

Se demuestra que a pesar de su estructura simple, el grafo planteado describe las relaciones usuario-tweet lo suficientemente bien, representando adecuadamente las cadenas de re-tweet y múltiples retweets realizados por el mismo usuario y que TURank evalúa a los usuarios que no son seguidos por muchos usuarios, pero sus tweets son re-tweeteados muchas veces, con una mayor posición en el *ranking*. Del mismo modo, los usuarios cuyos tweets no se re-tweetean, incluso si tienen un gran número de seguidores, se les asigna una peor la posición en el *ranking*. Por último, a aquellos usuarios en los cuales la mayoría de sus tweets son conversaciones, son evaluados como completamente inútiles por el algoritmo. Entregando de esta forma un eficaz sistema de puntuación para evaluar la autoridad de los usuarios de Twitter.

3.1.2. Geolocalización de usuarios

La geolocalización en Twitter no es una función muy utilizada por los usuarios según indicó Cheng en [8] tras el análisis de una muestra aleatoria de más de 1 millón de usuarios de Twitter, en la cual sólo el 21 % ha señalado el campo “ubicación” del usuario como un nombre granular de una ciudad (por ejemplo: Los Angeles, CA) y sólo 5 % ha señalado una ubicación tan granular como coordenadas de latitud / longitud (por ejemplo: “29.3712 , 95.2104”), el resto son demasiado generales (por ejemplo: California o España), no indican nada o uno sin sentido (por ejemplo: país de las maravillas). Como medio complementario para esta función, Twitter cuenta con la función de etiquetas geográficas ahora asociada a cada tweet. Pero a similitud del caso anterior, se observa que menos del 0,42 % de todos los tweets realmente utilizan esta funcionalidad.

McGee en [51] investiga la relación entre la fuerza del vínculo social entre un par de usuarios y la distancia entre dicho par con un conjunto de 6 millones de usuarios geolocalizados. En este estudio se observa que en una distribución bimodal en Twitter, con un pico de 10 kilómetros de las personas que viven cerca, y otro pico alrededor de 6.430 kilómetros, lo que valida el uso de Twitter tanto como una red social (con amigos geográficamente cercanos) y una plataforma de medios sociales (con conexiones muy distantes). También se observa que los usuarios con mayor fuerza en su vínculo (la amistad recíproca) tienen más probabilidades de vivir cerca entre sí que los usuarios con vínculos débiles.

Cheng por su parte, propone un marco de trabajo para la localización de usuario, que sea:

- Generalizable a través de las redes sociales.
- Robusto ante el ruido propio de los tweets.
- Confiable y preciso.
- Que trabaje únicamente con datos de dominio público por parte del usuario sin necesidad de análisis de otros datos de privacidad sensible (IP, usuario/clave, etc.).

El marco de trabajo propuesto se basa en la noción que los tweets de un usuario incluyen algunos contenidos de ubicación específica, como nombres específicos de lugares o palabras que se refieren a ciertos lugares además de otras denominaciones locales para los mismos (por ejemplo: "Valpo" para referirse a Valparaíso), y que con dicha información es posible cubrir la falta de una geolocalización para los usuarios.

La estimación de la localización basándose en el contenido de Twitter es una tarea difícil puesto que las actualizaciones de estado de Twitter son inherentemente ruidosas, a menudo con expresiones coloquiales y vocabulario no estándar. No es obvio que existan señales claras de ubicación incrustados en los tweets de algún usuario en lo absoluto. Un usuario puede tener intereses que abarcan múltiples ubicaciones y un usuario puede tener más de una ubicación física.

Este estudio acotado a los usuarios dentro del territorio continental de Estados Unidos. El primer filtro realizado se realiza con las ubicaciones del tipo: 'NombreCiudad', 'NombreCiudad, NombreEstado', 'NombreCiudad, AbreviaciónEstado' (considerando todas las ciudades válidas que figuran en el censo 2000 en EEUU). Para los casos en que existían dos ciudades con el mismo nombre, para determinar la ambigüedad, sólo se tienen en cuenta si poseen información adicional respecto al estado en el cual se ubican. Tras aplicar este filtro se encontró que sólo el 12 % de los usuarios de la muestra fueron identificados.

La segunda metodología implementada busca determinar la ubicación de un usuario en base a la contenido de los tweets de un usuario, caracterizándola por la distribución probabilista de sus palabras. Los resultados obtenidos con este enfoque es que sólo el 10.12 % de los usuarios pueden ser localizados a 160 kilómetros de sus ubicaciones reales (con una distancia de error promedio de 2853 kilómetros) lo que permite concluir que este enfoque no permite obtener resultados de calidad debido a:

- La mayoría de las palabras se distribuyen de manera compatible con la población a través de las diferentes ciudades, lo que significa que la mayoría de las palabras ofrecen muy poco potencial para distinguir la ubicación de un usuario.
- La mayoría de las ciudades, sobre todo con una población pequeña, tiene un conjunto escaso de palabras en sus tweets, lo que significa que la distribución por ciudades de palabras para estas ciudades está subespecificada, lo que conduce a grandes errores de estimación.

La tercera metodología determina la posición geográfica mediante la identificación de palabras locales en los tweets de los usuarios, basada en la noción que existen palabras que son más utilizadas en un ámbito geográfico que otros. Dichas palabras características son conocidas como palabras "locales" atribuidas a un territorio específico, si éstas son aisladas son capaces de distinguir a los usuarios situados en una ciudad y no en otra. El trabajo concluye que existe una gran posibilidad de clasificación considerando y aislando estas palabras debido a su potencial y establece un método para su tratamiento.

Las metodologías anteriores solo utilizaban como recurso de información los tweets, también en este trabajo se intenta obtener información adicional para localizar una persona basado en las relaciones entre usuarios en la red social. El vecindario de un usuario se determina mediante un vecindario difuso. Considerando a dos usuarios con una relación fuerte, cuando ambos usuarios se siguen mutuamente.

Finalmente con la combinación de las metodologías se logra una metodología de posicionamiento geográfico basado en la información de Twitter, el análisis lingüístico de sus tweets y el análisis de sus relaciones sociales con una precisión de 54.26 % y una distancia de error promedio de 760 kilómetros.

En [35] desarrollan una técnica de posicionamiento geográfico con usuarios en territorio Chileno, mediante un modelo de espacio vectorial que dado el contenido de un tweet permite a través de un clasificador establecer su ubicación geográfica, presentando un enfoque diferente a los trabajos revisados previamente basados en las características del usuario.

Mediante un clasificador TD-IDF y asumiendo que existen hashtags locales que se refieren a una ubicación concreta, es posible deducir la ubicación de un tweet con las palabras que co-ocurren junto a estos hashtags. Por lo cual la asociación entre los hashtags y las palabras claves

locales identificadas se utilizan para geolocalizar usuarios utilizando un clasificador construido sobre modelos de lenguaje.

Finalmente los resultados concluyen que los participantes en Twitter son representativos de la población física, es decir también responden al centralismo del país.

3.2. Herramientas y plataformas relacionadas

3.2.1. Geofeedia

Geofeedia [33] es un servicio pionero en el control de origen de los medios de comunicación social mediante geolocalización, permite el monitoreo y realizar análisis a datos provenientes de las redes sociales, complementando la tradicional búsqueda de *keywords* que típicamente pierden información de la actividad social relacionada, ayudando a reducir el desorden de los medios sociales en tiempo real, mejorando los descubrimientos escondidos de la información geolocalizada. Geofeedia realiza búsquedas en las redes sociales: Twitter, Instagram, YouTube, Picasa, y Flickr.

Posee una interfaz intuitiva y fácil de usar además de un conjunto de herramientas para facilitar las búsquedas. Geofeedia permite obtener stream de datos de redes sociales provenientes de una zona geográfica en específico, mediante la mención textual del lugar o la delimitación de una región en un mapa interactivo.

Los diversos feeds se pueden categorizar y olvidar usuarios (no recibir feed de dicho usuario). Los resultados de la búsqueda se pueden filtrar por fecha, por palabras claves, por tipo de medio social y por autor, entre muchos otros. La plataforma además permite variadas herramientas de análisis como identificar la tendencia de palabras claves, la actividad basada en el tiempo, frases influyentes, las fuentes de los medios sociales.

Los datos se pueden exportar a través de la API con todas las funciones de Geofeedia en formatos CSV ATOM, GeoRSS, JSON, KML.

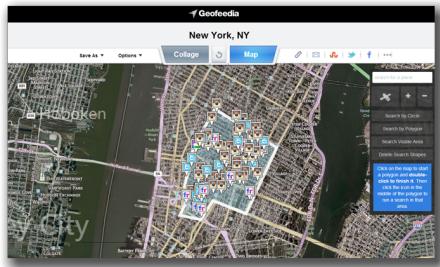


Figura 3.1: Imagen del mapa interactivo de geofeedia donde en una zona delimitada por el usuario se reciben todos los feeds de los medios sociales.

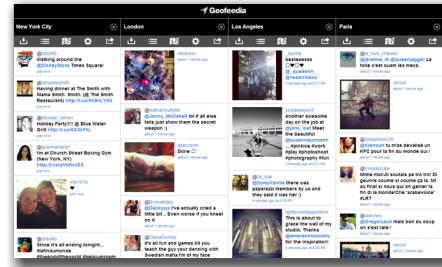


Figura 3.2: Imagen del panel de noticias de geofeedia, donde cada columna recoge los feeds para ubicaciones geográficas distintas.

3.2.2. Paper.li

Paper.li [53] es una creación de la startup suiza Small-Rivers con sede en Lausanne. Es una herramienta que permite generar una especie de periódico personalizado en base a las publicaciones en redes sociales de los contactos (en dichas redes) del usuario. Actualmente Paper.li permite el acceso a fuentes como Twitter, Facebook, Google, YouTube y los canales RSS para la recolección de información.

Permite fáciles combinaciones de las distintas fuentes de contenido, la priorización temática, la aplicación de filtros de contenidos, configurar las fuentes de información y personalizar el aspecto del periódico personal.

El funcionamiento de Paper.li se basa en el hecho de que muchos usuarios de Twitter confían más en el criterio de selección de sus redes sociales para identificar enlaces a noticias importantes que en las compilaciones que realizan los diarios tradicionales. Paper.li recoge los enlaces a noticias, fotos y vídeos de una cuenta en Twitter y otras redes sociales; realiza una selección de estos vínculos (a partir de un análisis realizado con “herramientas de análisis semántico de texto”), crea una página diaria editada con aspecto a una página de los periódicos tradicionales que vemos en Internet, donde los enlaces y el contenido aparecen divididos en secciones contextuales.

Los creadores de Paper.li (mediante el blog oficial del medio) han declarado que la motivación inicial para crear este medio “no es por la sobrecarga de información es por la ausencia de un filtro”. Es importante señalar que Paper.li no cumple exactamente con la tarea de entregar las noticias de forma urgente debido a su intervalo de actualización diario, Paper.li más bien,

reporta el eco de ellas en las redes sociales, la reproducción de contenidos más frecuentes, los tópicos más consultados o lo más recomendados.

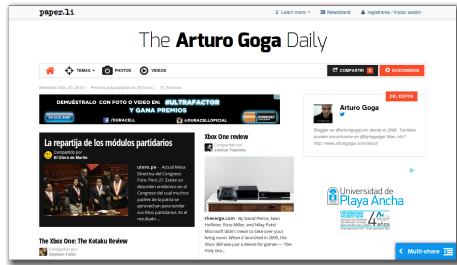


Figura 3.3: Vista principal del periódico Paper.li de un usuario de la plataforma

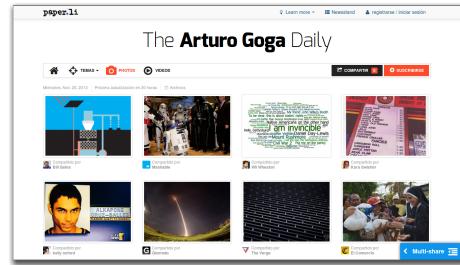


Figura 3.4: Vista de fotografías del periódico Paper.li de un usuario de la plataforma

3.2.3. The Tweeted Times

The Tweeted Times [71] es una aplicación web lanzada en 2010 y desarrollada por Flipboard, Inc. permite generar en base a los *feeds* recibidos en la cuenta de twitter, una especie de periódico de noticias con las temáticas más importantes, de forma similar a Paper.li 3.2.2 pero sólo enfocado en Twitter y con la potencialidad de que se actualiza cada una hora.

The Tweeted Times posee otras potencialidades como su capacidad de personalización en la selección de los contenidos. Además de la cuenta del mismo usuario, es posible crear periódicos con listas particulares de Twitter, con perfiles de usuarios, y hasta de una búsqueda de un hashtag o de usuarios.

The Tweeted Times usa un algoritmo particular para poder organizar las publicaciones de acuerdo con su correspondiente relevancia. El algoritmo se basa principalmente en la repercusión de dichas publicaciones en Twitter (cantidad de retweet, número de marcaciones como publicación favorita).

A cada enlace compartido se le asigna el valor de un *ranking* relacionado con la cantidad de replicaciones por amigos y por amigos de amigos y menciona por cual de los contactos fue aportado el enlace. Permite explorar los periódicos de otros contactos en Twitter y también periódicos de otros usuarios destacados de la plataforma.

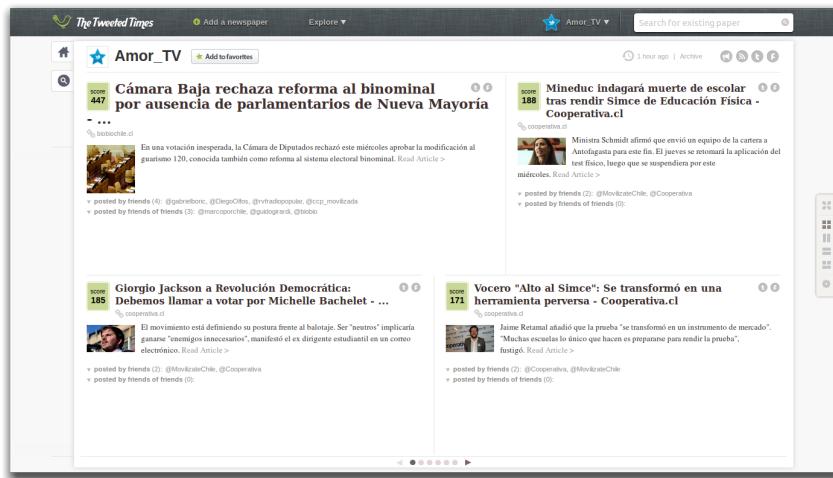


Figura 3.5: Vista principal de un periodico creado en The Tweeted Times

3.2.4. FlipBoard

Flipboard [25] es una aplicación lanzada en julio de 2010 que permite reunir las noticias del mundo y las novedades de las redes sociales en una revista diseñada para dispositivos móviles. El usuario puede elegir algunas temáticas y Flipboard crea una revista digital, en la cual se pueden “hojear” las noticias de interés, así como las historias y fotos que los contactos del usuario comparten. Flipboard posee además una función que permite guardar artículos para verlos posteriormente o recopilarlos en “revistas” propias de Flipboard.

Flipboard actualmente permite la conexión con 12 redes sociales entre las que se incluyen Twitter, Facebook, Instagram, Google+, YouTube, Google Reader, LinkedIn, Flickr, 500px, Sina Weibo y Renren, en las cuales es posible realizar búsquedas mediante temas, hashtag, blogs y personas. Hoy en día existen 15 diferentes versiones de Flipboard entre países de América, Europa y Asia.

Flipboard también pone a disposición de sus usuarios contenidos seleccionados donde se incluyen las revistas y blog recomendados, fotografías y secciones especiales a noticias del día y temas de interés.

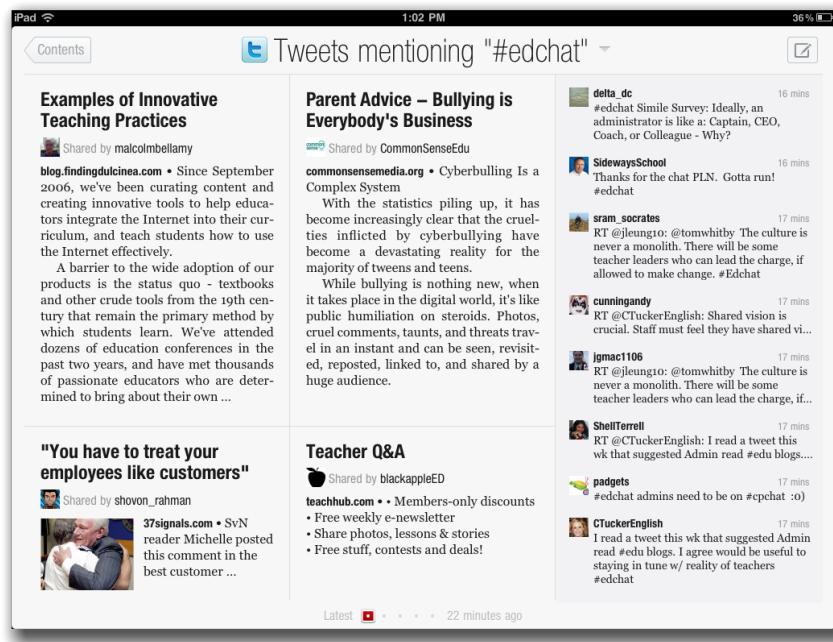


Figura 3.6: Vista del formato de revista en Flipboard para visualizar los feeds de Twitter, donde se resaltan los tweet que han tenido mayor repercusión en la red y los enlaces compartidos

3.2.5. Summify

Summify[69] consistía en un servicio gratuito que permitía informarse de la actualidad por medio de recopilaciones diarias de las noticias que más circulaban entre el círculo de contactos de Twitter y Facebook, la cuenta de Google Reader o feeds RSS del usuario. Summify cesó sus funciones el 22 de junio de 2012 por decisión de Twitter a solo cinco meses de comprarlo. Summify entregaba su servicio de distintas formas: mediante la web para revisar las recopilaciones, vía correo electrónico, mediante la notificación por DM de Twitter cuando existe una nueva disponible y mediante la aplicación móvil.

La principal motivación para esta aplicación era expresada mediante su video promocional: En un inicio mantenerse al tanto de las actividades sociales es muy gratificante, pero luego, fácilmente, se puede convertir en una actividad agobiante y poco productiva por la gran cantidad de contactos y las diversas redes sociales existentes. Summify se presentaba como una alternativa productiva para mantener satisfactoriamente los vínculos sociales en la red.

Summify permitía la configuración de la cantidad de recopilación se deseaba recibir el

usuario (de una a cuatro diarias) y a la hora de recopilación de la primera de éstas (las restantes se realizaban en consecución en horas), era posible especificar la cantidad de artículos que poseían dichas recopilaciones (de dos a quince artículos), además de la privacidad de dichas recopilaciones (públicas o privadas). Su algoritmo se encargaba de recopilar los enlaces a artículos que aparecen en las cuentas del usuario en Twitter y Facebook y los filtraba en base a cuanto han sido compartidos, marcados con “Me gusta” o re-tweeteados, prestando atención a las personas con las que el usuario interactua más. También el algoritmo consideraba que tanto se ha compartido un enlace a nivel global, pero según indicaban en su blog de documentación, tenía menos relevancia que las interacciones del usuario con sus contactos. Otra cualidad del algoritmo que se utilizaba es que aprendía en base al uso dado por el usuario, analizando las tendencias y preferencias de uso, observando si se pinchaban más historias de un usuario en particular, en las de un dominio en concreto, en las que contienen unas palabras claves concretas dentro del título o cuál es la fuente de feeds más utilizada.

Cada recopilación poseía su propia página individual y mostraba las noticias seleccionadas una debajo de la otra. Junto a ellas, se visualizaban algunos de los usuarios que las habían compartido y al posicionar el puntero por encima de los avatares, el mensaje en el cual compartieron el artículo. Todas las recopilaciones se conservaban y eran posible acceder a ellas desde el perfil del usuario.

En el caso de que algunos de los contactos del usuario en Twitter y Facebook también estén usando Summify, era posible acceder a sus recopilaciones (en el caso de que éstas fuesen públicas). Era posible también navegar por los perfiles y recopilaciones de los usuarios que se siguen, que siguen al propio usuario y, además, de los que se considera que influencian al usuario y a los que influencian al usuario (es decir, que aparecen más veces en los resúmenes del usuario o en los que aparecemos más veces el propio usuario).

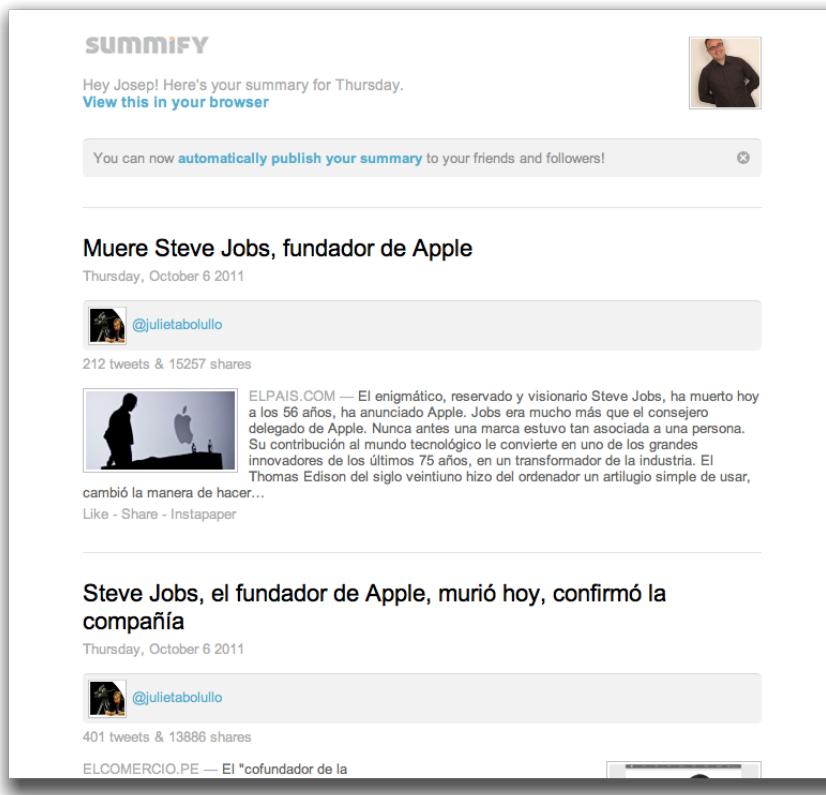


Figura 3.7: Vista de la recopilación de Summify

3.2.6. TwitterFall

Twitterfall [?] es un cliente web basado para Twitter que permite realizar búsquedas avanzadas y seguimiento de temas de tendencia. Fue construida el 2009 por Tom Brearley y David Somers, dos estudiantes de ciencias de la computación de 19 años en la Universidad de York y después de tres años de funcionamiento ha sido denominado como “el google para la tweetosfera”.

El nombre de la plataforma se relaciona con la forma de su interfaz, ya que se presenta como una cascada de tweets que van cayendo hacia abajo de la pantalla a medida que existen nuevas actualizaciones en tiempo real tal como se puede ver en la figura 3.8. La velocidad de este flujo puede ser personalizada mediante un simple ajuste.

Las potencialidades de TwitterFall radican en sus herramientas de búsqueda, entre las cuales se encuentran:

- Realizar búsquedas mediante palabras claves o trendic topic.

- Focalizar búsquedas geolocalizadamente.
- Se puede realizar filtrados por idioma y excluir re-tweet.
- Permite almacenar las búsquedas favoritas.

TwitterFall permite también iniciar sesión en Twitter desde su interfaz y utilizarlo como cliente regular de Twitter y utilizarlo para responder mensajes, realizar re-tweet, seguir a nuevas personas, etc. otorga además interesantes características como pre-visualizaciones en ventanas flotantes sobre el enlace contenidos en tweets (sin necesidad de ser redirido a ellos), o tras posicionar el puntero sobre un tweet se obtiene información sobre su ubicación y sus seguidores.



Figura 3.8: Vista de la recopilación de Twitterfall

3.2.7. Storyful

Fundada en Dublín con sólo tres empleados, Storyful [76] tiene su razón de ser en idear una manera de administrar las enormes cantidades de contenido que se comparten en las redes sociales por millones de usuarios en todo el planeta, aplicando procesos y tecnología eficaz para ayudar a filtrar “noticias de ruido”.

A finales del 2011 Storyful formó una alianza con YouTube para encontrar y la validar los videos que surgieron a partir de las protestas crecientes en Egipto.

Storyful pretende dotar de herramientas a todos los medios de prensa que las demanden, tal como declaran mediante su página web “Desde 2010, Storyful ha estado construyendo y perfeccionando la primera sala de prensa verdaderamente social. Hemos perfeccionado nuestras técnicas, herramientas y servicios, en colaboración con algunas de las marcas más importantes

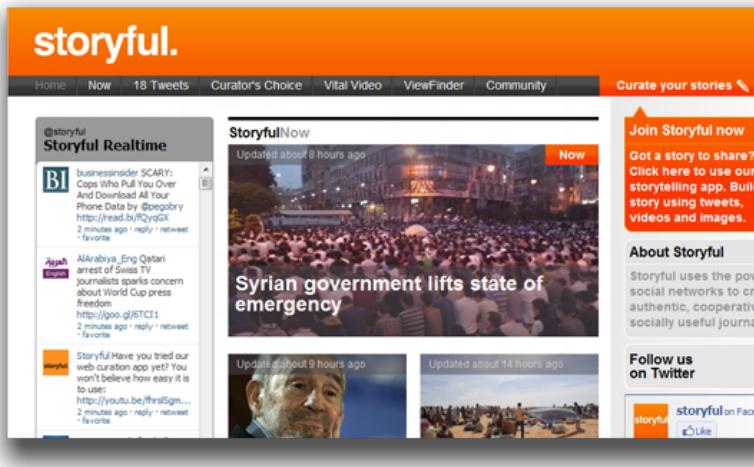


Figura 3.9: Vista principal de Storyful

de noticias en el mundo, incluyendo ABC News, Reuters y el New York Times, y plataformas sociales como YouTube.

Storyful ha desarrollado lo que llaman como “algoritmo humano” que consiste en un conjunto de procesos algorítmicos y humanos que posee gran aceptación hoy en el mercado de las noticias obtenidas de medios sociales.

Para los nuevos medios sociales Storyful plantea que los periodistas deben pensar en su papel en un mundo cambiante, comenzando a utilizar nuevas herramientas inteligentes y rápida, complementándolas con los conocimientos profesionales y de oficio de los periodistas. Separando el proceso en tres etapas: descubrimiento de hechos noticiosos, verificación y entrega o presentación de los contenidos.

Storyful asegura que la única manera de verificar una información y restringir correctamente su ruido es unirse a la conversación sostenida en los medios sociales desde donde surgen los hechos noticiosos aprovechando las oportunidades que ofrecen hoy en día las tecnologías actuales, pero no sólo escuchar en ellos, sino participar directamente, de forma abierta y honesta con las voces más cercanamente relacionadas con el hecho. Esta es una clave del algoritmo humano ”La mayoría de las organizaciones de noticias se busque en observadores pasivos. Pero eso no es suficiente. La única manera de desbloquear el poder de la Algoritmo humano es ser una parte de ella”³.

³<http://storyful.com/about/>

Señalan que el valor de la información al relacionarse en los grupos de redes sociales es increíble, apasionante y de rápida evolución, pudiendo explicitar relaciones con otros hechos noticiosos aparentemente invisibles, que finalmente enriquecen la noticia y la completan, permitiendo contacto y conversaciones abiertas, informadas y sinceras con los lugareños del mismo sector o con académicos de la zona, con información que de otro modo difícilmente fuese recolectada.

Storyful plantea que fue diseñado para “vivir dentro de las comunidades de medios sociales, no para observar desde una distancia segura”. Para lo cual cuenta con jueces en terreno que establecen vínculos sociales y contactos con personas del entorno para acreditar desde estas relaciones, la veracidad de la información, complementando el proceso de verificación establecido para un hecho en específico.

El proceso de verificación es la piedra angular del proceso, puesto que muchas veces en las redes sociales se hacen circular deliberadamente informaciones incorrectas, generando desinformación y percepciones erradas. En dicho proceso aplican las siguientes preguntas para generar un cierto nivel de fidelidad de la información aportada por un cargador en Youtube:

- ¿Dónde se ha registrado esta cuenta y en qué se ha basado el autor para emitir su juicio de la historia?
- ¿Existen otras cuentas en Twitter, Facebook, blogs o sitios web relacionados a este autor ? ¿Qué información tienen para indicar la ubicación reciente, la actividad, la fiabilidad y/o el sesgo?
- ¿Hace cuánto tiempo existen esas cuentas? ¿Qué tan activas son?
- ¿La narración del contenido se escribe en jerga o dialecto identificable en la narración de la publicación?
- ¿Podemos encontrar información WHOIS⁴ para algún sitio web afiliado ?
- ¿El cargador aparece en los directorios locales ? ¿Qué indican sus círculos sociales en línea que están cerca de esta historia / ubicación?

⁴Es un protocolo de petición y respuesta ampliamente utilizado en bases de datos para el registro de usuarios o asignaciones de algún recurso de Internet, como el nombre del dominio o el bloque de dirección IP

- ¿Tiende el cargador a 'copiar' contenidos de otras cuentas de Youtube , o carga únicamente contenido generado por el usuario?
- ¿Los videos de la cuenta del cargador tienen una calidad consistente?
- ¿La descripción del video es consistente y señalan un lugar específico ? ¿Corresponde la fecha? ¿Tiene la extensión de archivo como AVI o MP4 en el título del video?
- ¿Storyful esta familiarizado con esta cuenta? ¿su contenido ha sido fiable en el pasado?

Storyful comparte sus conocimientos y técnicas de forma abierta desde el blog de su plataforma los cuales se revisan a continuación debido a su posible aplicabilidad como directrices en este trabajo:

- **Hacer listas: Cómo afinar sus antenas Twitter [67]**

Una herramienta fundamental para el funcionamiento de Storyful, son las listas temáticas de Twitter. Debido al amplio flujo de información en la era de los medios sociales, las listas funcionan como el filtro perfecto de noticias en vivo, permitiendo afinar el lugar de observación, un objeto o un evento y silenciar algunas conversaciones no útiles que tienen lugar simultáneamente.

Maneján simultáneamente cientos de listas de Twitter que cubren todos los países del mundo y todos los temas imaginables es complejo.

Para la administración de estas listas se utiliza la herramienta *Tweetdeck*[72], donde a cada terminal se le cargan cerca de 20 a 30 listas para ser visualizadas por un periodista. *Tweetdeck* es utilizado (a pesar de sus limitantes en el número de listas que permite por usuario) por las facilidades de gestión que aporta como la agrupación y orden dinámico de las listas por temáticas similares, clasificación e inclusión de nuevos elementos (como usuarios o tópicos) mediante mecanismos interactivos.

- **Verificación de videos [68]**

- Revisión de la historia y la ubicación del cargador para ver si ha compartido contenido útil y creíble en el pasado.

- Uso de las imágenes Google Street View para ayudar a verificar los lugares en un video.
- Consulta de otras fuentes de noticias o contenido de usuario validado para confirmar eventos ocurridos en el video.
- Examinar características claves del video como el clima y el paisaje de fondo para ver si coinciden con los hechos conocidos en terreno.
- Traducción de toda palabra que sale en el vídeo a modo de contexto adicional.
- Supervisar el tráfico de las redes sociales para ver quién está compartiendo el contenido y qué preguntas se les pide al respecto.
- Desarrollar y mantener relaciones con las personas dentro de la comunidad alrededor de la historia.

3.2.8. Wikipulse

Wikipulse es un medio de prensa [31] basado en el hecho de que la propia comunidad de editores de Wikipedia muchas veces logran actualizar este portal y anunciar un hecho, mucho antes que los medios de prensa convencionales. Wikipulse recoge las más recientes ediciones y las presenta en un formato ordenado y simple.

Existen estudios [70] que verifican que existen muchos editores en Wikipedia que se encargan de actualizar lo más pronto posible diversos los artículos, motivados por la mejora en su reputación como publicadores que esto implica, dichas actualizaciones llegan incluso a hacerse públicas en un tiempo promedio de dos horas inferior al reporte de noticieros como CNN o *Reuters online* [3].

Wikipulse realiza su trabajo en tres principales tareas:

1. Encontrar artículos relevantes en Wikipedia
2. Reformar los artículos en formato de noticias
3. Mostrarlos en una página web

El núcleo como tal de Wikipulse esta compuesto por dos componentes: Wikipedia actúa como fuente principal de artículos de noticias mediante artículos editados o publicados por al

comunidad de Wikipedia, y la segunda componente esta compuesta por la generación de noticias de Wikipulse de forma automática obtenidos desde Wikipedia agrupados en las distintas categorías (eventos corrientes, deporte, etc.).

Para el criterio de selección además del criterio obvio de los artículos editados más recientemente en Wikipedia, se emplea un algoritmo planteado en un trabajo reciente en [30] se observa que en la edición de un artículo se genera una red basada en vínculos de “editores-compartidos”: dos artículos diferentes tienen un vínculo entre ellos si el mismo editor modifica ambos

La selección de los autores “correctos” garantiza artículos de interés periodístico por el mantenimiento de una lista cada vez mayor de editores frecuentes que se centran en el artículo-tipo de prensa. La comprobación de si los hechos del artículo son verídicos ocurre de forma gratuita a través del principio de los muchos globos oculares de Wikipedia (muchos usuarios ven y corroboran si la información es verídica, si no lo es, la editan).

En el segundo paso se formatean los artículos de Wikipedia a un estilo más periodístico, usando técnicas automática de generación abstracta. El paso final consiste en mostrar las artículos en un formato periódico de manera online.

El sistema se encuentra conformado por cuatro capas: Wikipedia, La capa de extracción, la capa de selección de noticias y la página de presentación.

Algoritmo de selección de noticias

El algoritmo se basa en tres grandes etapas:

1. **Creación del Conjunto de trabajo:** El algoritmo es ejecutado periódicamente y en cada ejecución procesa un conjunto de noticias, las cuales son llamadas *conjunto de trabajo*. Este *conjunto de trabajo* contiene el conjunto de páginas recientemente editadas en Wikipedia. Cada una de estas páginas es pasada posteriormente por las otras componentes del algoritmo. En primer lugar la página de metadatos, el título y la totalidad de los autores y todas sus categorías son guardadas en una base de datos de un grafo de autores, permitiendo posteriormente a otras partes del algoritmo realizar consultas sobre esta información. El proceso de almacenamiento asegura que la base de datos es actualizada y expandida con cada ejecución del algoritmo aun cuando no existen nuevas noticias son generadas.

2. Selección de Noticias: El proceso de selección busca páginas de Wikipedia que podrían llegar a ser noticias posteriormente. Esto es mediante la generación. Esto se realiza mediante la generación de un ranking r por cada página, éstos son comparados con un valor de umbral t . Cada rank esta basado en consultas a la base de datos, ponderadas por un peso w que especifica la importancia de ese rank en el resultado final. Una página genera una noticia cuando se cumple que:

$$\sum_{i=1}^n r_i \cdot w_i > t_i$$

Cada rank diferente que es generado por cada página son explicados a continuación. La mayor parte de ranks son ratios computados en orden para tener un conjunto predecible de números como resultado.

- Autor-con-Noticia el rank Autor-con-noticia aumenta la importancia de las páginas autores generadores-de-noticias(autores que ya han generado noticias). Y se calcula como la cantidad de autores que editan la página p partido por todos los autores generadores de noticias.
- Autores-comunes Calcula la popularidad de una página p mediante sus autores. Calculando el ratio de todos los autores editando la pagina p por todos los autores de la base de datos.
- Dominio-experto El rank dominio-experto identifica páginas importantes mediante el conteo del conjunto de expertos que editan la página. Un experto del dominio es aquel autor en Wikipedia que edita páginas en la misma categoría que la página que se esta evaluando.
- Cambios-recientes Este rank mide la actividad de edición en la página en comparación con todas las otras páginas del conjunto de trabajo.
- Relevancia La relevancia utiliza una evaluación de un webservice externo para determinar la popularidad de una página de Wikipedia.

3. Creación de noticias: El proceso de creación de noticias crea noticias a partir de las páginas. Esto se hace mediante el análisis y la agregación de las ediciones de una página en una gran bloque de texto, que se envía a continuación a un webservice llamado

summry.com que las resume en un artículo de noticias cortas. Este artículo de noticia es guardado en la base de datos, y se presenta a través de la interfaz web. Para agregar una edición, esta es elegida en base a tres criterios:

- La edición debe ser hecha por un editor top.
- La edición debe ser hecha por un experto del dominio.
- La edición es más larga que 50 caracteres.

Un editor top es aquel autor que está dentro de los “editores top” lista que incluye los autores y expertos de dominio que tienen más ediciones para una categoría dada.

Capa de extracción

Es la capa encargada de mediante una conexión a la API de Wikipedia permite el acceso a las páginas modificadas y sus datos respectivos. Esta capa utiliza algunas herramientas externas:

- *summry.com* una API abierta que permite resumir textos de páginas con un número excesivo de oraciones. (Disponible en <http://summry.com/api>).
- *stats.grok.se* un subproyecto de Wikipedia el cual mantiene estadísticas de las diversas páginas de Wikipedia.

Análisis de los datos

El análisis de datos se realizó con intenciones de medir el rendimiento del seleccionador de noticias, de esta forma se decidió compararlo con algunas fuentes de información tradicionales como (CNN, Reuters, etc.) La idea principal es comprobar si las noticias seleccionadas por Wikipulse también fueron registrados como noticias por los medios de comunicación convencionales (dentro de un límite determinado de tiempo). Para dicha medición se utilizaron los siguientes criterios:

- Precisión y relevancia Medida del solapamiento entre las noticias de Wikipulse y las reporteadas por los medios tradicionales en algún tiempo dado.
- Frescura/rapidez Criterio que mide el tiempo relativo respecto al solapamiento de las noticias entre Wikipulse y los medios convencionales. Este indica si Wikipulse es más rápido o más lento en términos de reporte.

Para comparar los resultados de Wikipulse se compararon los reportes con un medio convencional mediante el acceso a su RSS y así obtener el momento de inicio en la difusión de la noticia. Para comparar si se trataban del mismo contenido, se realizó un proceso de *match* entre ambos recursos de la siguiente manera: Las historias individuales obtenidas de ambos *feeds* son procesados para obtener una lista de palabras. Los distintos artículos de ambos feed son comparados mediante el *match* de las palabras claves. El cuantificador se calcula mediante una fracción entre el número total de match encontrados dividido por el número total de *keywords* en cada uno de los artículos. Empíricamente se encontró que un cuantificador superior a 0.33 generalmente indicaba un *match* fuerte.

Capítulo 4

Definición de la solución

Existe un creciente sentimiento de no seguridad por parte de la ciudadanía y múltiples teorías de investigadores y científicos respecto a la repercusiones reales generadas por el proceso de *gatekeeping* en las noticias publicadas por los medios de prensa tradicionales.

La irrupción de Internet y las redes sociales en la vida cotidiana han modificado variados esquemas de comunicación y formas de relación entre las personas, entre ellos, los medios de prensa y la forma que tienen los ciudadanos para informarse de los eventos que ocurren en su entorno y/o que concitan su interés. Tambini en¹ señala “El papel de *gatekeeper* de los medios de prensa tradicionales se ha debilitado y las personas construyen su propia estructura editorial eligiendo qué siguen en los medios de comunicación basados en la recomendación. Las redes sociales también difieren en el sentido de que tienen un compromiso con la universalidad y apertura, que es más importante para ellos que los periódicos estén ahí”

Twitter debido a las múltiples características que posee referente a su capacidad de propagación, acceso y espontaneidad se presenta como una importante red social que hoy revoluciona la industria periodística mundial, obligando a las grandes cadenas a integrarles dentro de su proceso periodístico. Hughes, un corresponsal muy influyente y famoso de la BBC, dice: “Los medios sociales permiten que consiga mucho más acerca de la historia. Hay periodistas y otras personas en el mismo lugar del suceso que reportan los eventos en tiempo real, de manera que cuando una historia en realidad aparece en el cable, a menudo ya la he visto través de los medios sociales. Hughes declara, que al igual que una gran cantidad de periodistas, recolecta

¹<http://wallblog.co.uk/2013/03/05/how-twitter-won-the-social-media-battle-for-journalism/#ixzz2kBQ2RFcQ>

inicialmente una gran cantidad de noticias desde Twitter. Otros informes dicen que la cifra de periodistas recibiendo historias y eventos de Twitter es de aproximadamente 50 %. Hughes dice que el 80 % de su recopilación de noticias la obtiene en Twitter y sólo el 20 % de otras fuentes.²

Considerando el interés creciente de adquirir información de eventos noticiosos desde sus fuentes directas evitando los procesos de *gatekeeper* sumado a las potencialidades que posee Twitter referente a esta misma temática se busca diseñar e implementar una herramienta que permita comunicar los reportes en Twitter de un determinado evento, privilegiando aquellos reportes con menor cantidad de intermediadores y posibles *gatekeeper*, siendo estos, los observadores y reportadores directos del suceso.

4.1. Objetivos

4.1.1. Objetivo principal

- Desarrollar un algoritmo computacional que permita recoger tweets que reporten un evento, priorizando los tweets geolocalizados cercano al lugar de ocurrencia del evento, para generar un relato temporal referente a dicho evento que será presentado a otras personas mediante una interfaz web.

4.1.2. Objetivos Secundarios

- Analizar trabajos previos y herramientas creadas con anterioridad para encontrar la forma adecuada y más conveniente de proceder a la construcción de esta herramienta.
- Dotar al público interesado en informarse sobre eventos noticiosos de una herramienta de reporte de eventos que minimiza el filtrado y tratamiento editorial de contenidos.
- Desarrollar una interfaz web que sea clara y fácil de usar por el usuario.

²http://www3.ebu.ch/files/live/sites/ebu/files/Knowledge/Media%20Strategy/ebu_polis_report1.pdf

Capítulo 5

Propuesta

5.1. Arquitectura de la solución

La Arquitectura de la solución se compone de los siguientes macroprocesos conceptuales:

- Captura de datos: Captación de usuarios y captación de tweets.
- Almacenamiento.
- Procesamiento de tweets: Extracción de datos e Indexación de datos.
- Presentación de los datos.

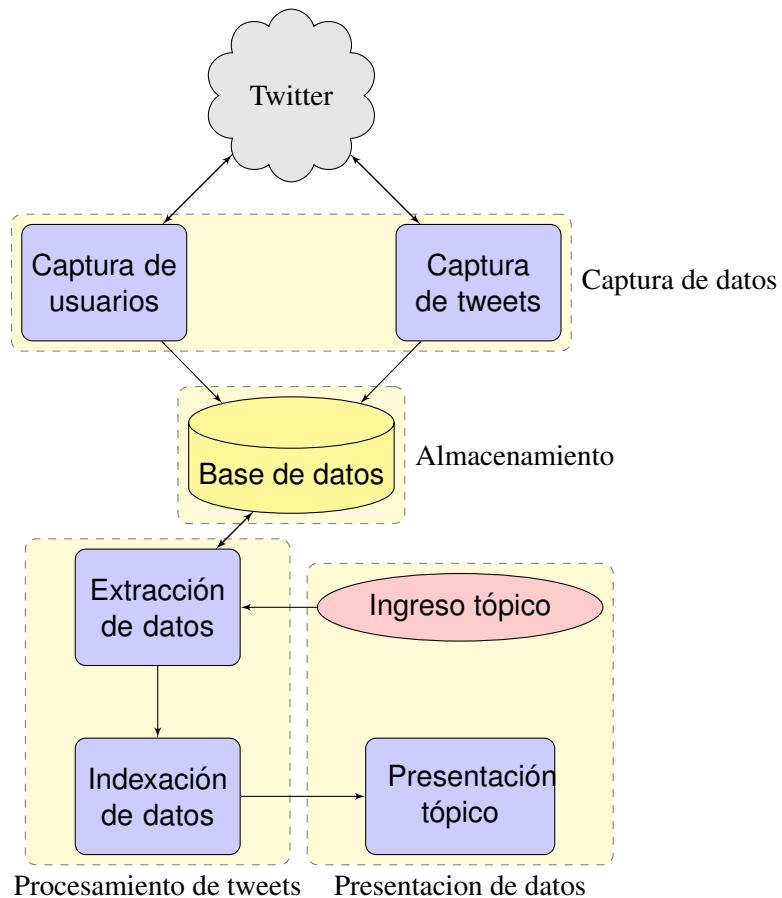


Figura 5.1: Diagrama conceptual de la arquitectura de la solución

5.2. Plataformas y herramientas utilizadas

Las plataformas y herramientas utilizadas para desarrollar el prototipo fueron las siguientes:

■ Python

Python es un lenguaje interpretado de programación orientado a objetos. Incorpora módulos, excepciones, interpretación dinámica, tipos de datos dinámicos de muy alto nivel. Según su propia página web [28] *Python combina una notable potencia con una sintaxis muy clara. Tiene variadas interfaces para muchas llamadas de sistema y bibliotecas así como de sistemas de ventanas y es extensible en C o C++.* Otra cualidad notable de Python es que es portátil y se ejecuta en la gran mayoría de variantes de UNIX, Mac y

Windows.

Python cuenta actualmente con una gran comunidad que genera y mantiene una robusta documentación lo que lo vuelve más accesible al aprendizaje.

El lenguaje viene con una biblioteca estándar que cubre áreas como procesamiento de strings (expresiones regulares, unicode, diferencia de archivos), protocolos de internet (HTTP y FTP entre otros), ingeniería de software y las interfaces del sistema operativo (llamadas al sistema operativo y sistemas de archivos). Python cuenta también con una gran variedad de extesiones desarrolladas por terceros [29], algunas de las cuales fueron ocupadas en este trabajo.

■ MySQL

MySQL es un sistema de gestión de bases de datos relacionales, multihebra y multiusuario con más de seis millones de instalaciones. MySQL AB desarrolla MySQL como software libre en un esquema dual bajo licencia GNU GPL y licencia de pago para productos privativos. MySQL destaca por su gran adaptación a diferentes entornos de desarrollo permitiendo su interacción con los lenguajes de programación más utilizados como PHP, Perl, Python y JAVA.

MySQL posee las características distintivas de otros motores de bases de datos:

- Permite escoger entre múltiples motores de almacenamiento para cada tabla (entre los que se encuentra MyISAM, Merge, InnoDB, Memoryheap y muchos más).
- Permite la agrupación de transacciones para mejorar el número de transacciones por segundo.

Mysql es usado por muchos sitios web populares como Wikipedia, Google, Facebook, Twitter, Flickr y Youtube.

■ Pycharm

Pycharm [64] es una IDE utilizada para programar en Python. Esta IDE provee análisis de código, un *debugger* gráfico, una unidad integrada para pruebas e integración con sistemas de control de versión además de soporte web para desarrollar con el framework Django.

Algunas empresas que utilizan Pycharm son: Ebay, Groupon, Linkedin, Twitter, Spotify y HP. Pycharm cuenta con una licencia Profesional libre para proyectos de código libre y para fines educacionales, cuenta también con una edición *community*

■ Django

Django [26] es un framework de alto nivel de desarrollo web en Python que fomenta el desarrollo rápido y el diseño limpio y pragmático. Django es gratuito y de código abierto y se basa en el patrón de arquitectura MVC.

El objetivo principal de Django es facilitar el desarrollo de sitios web complejos con bases de datos. Django potencia la reutilización y conexión de componentes, el rápido desarrollo y el principio de no repetir código. Django ofrece también una herramienta administrativa opcional que mediante una interfaz web comúnmente utilizada por los administradores del sistema web para crear, modificar y leer los datos de la plataforma web en cuestión.

Las componentes de Django se basan principalmente en un modelo MVC, tratándose de un mapeador objeto-relacional que media entre los modelos de datos (definidos como clases de Python) y una base de datos relacional (llamado *modelo*), un sistema para el procesamiento de las peticiones HTTP con un sistema de plantillas web (llamados *visitas*) y una despachador basado en expresiones regulares de URL (llamado *Controlador*).

Django soporta oficialmente cuatro bases de datos backend: PostgreSQL, MySQL, SQLite y Oracle.

El framework también incluye:

- Un servidor web ligero y autónomo para el desarrollo y pruebas
- Un formulario de serialización y un sistema de validación el cual puede traducir entre los formularios HTML y los valores esperables para el almacenamiento en la base de datos
- Un sistema de plantillas que utiliza el concepto de herencia
- Un sistema de internacionalización, incluyendo traducciones de propios componentes de Django en una variedad de idiomas.

Algunos sitios conocidos que utilizan Django son: Pinterest, Instagram, Mozilla y The Washington Times y cuenta actualmente con una activa comunidad de decena de miles de usuarios y colaboradores en todo el mundo.

■ **Mysql Workbench**

MysqlWorkbench [12] es una herramienta visual de diseño de bases de datos que integra el desarrollo de SQL, administración, diseño, creación y mantenimiento de bases de datos en un único entorno de desarrollo integrado para MySQL.

MySQL Workbench proporciona herramientas visuales para crear, ejecutar, y optimizar consultas SQL. El editor de SQL proporciona color resaltado de sintaxis, autocompletado, la reutilización de fragmentos de SQL y el historial de ejecución de SQL. El panel de conexiones de base de datos permite a los desarrolladores para gestionar fácilmente las conexiones de base de datos estándar, incluyendo Tela MySQL. El Examinador de objetos proporciona acceso instantáneo a esquema y objetos de base de datos.

■ **Api de twitter**

Twitter mediante sus API [44] habilita para que desarrolladores puedan escribir y leer en Twitter. Para el acceso y manipulación de datos de Twitter existen dos API disponibles: Rest API y Streaming API. La Rest API permite crear nuevos tweets, conocer información sobre el autor de un tweets entre otras acciones relacionadas a tweets o usuarios particulares. La Streaming API entrega un flujo constante de información en tiempo real.

La Rest API de Twitter identifica aplicaciones mediante OAuth, posee una limitación de 150 solicitudes por hora cuya repercusión afecta directamente el desempeño de los distintos algoritmos de captura de datos. Debido a que no existe una necesidad de instantaneidad prioritaria y por el volumen acotado de datos a manejar en este trabajo se utilizó la Rest API de Twitter.

Para su utilización de manera simple se utilizo la librería Python Twitter [56] [10] que provee una interfaz en python para la API de Twitter.

■ **Librerías de Python utilizadas**

- **LXML**

LMXL [65] es un conjunto de herramientas que vincula las librerías C libxml2 y libxslt para su uso en Python, combinando la velocidad y la exhaustividad de las funciones para el análisis de XML de estas librerías con la simplicidad de Python. Implementa los siguientes protocolos: *XML 1.0, HTML 4, XML namespaces, XML Schema 1.0, XPath 1.0, XInclude 1.0, XSLT 1.0, EXSLT, XML catalogs, canonical XML, RelaxNG, xml:id, xml:base*.

Posee una basta documentación debido a que implementa ElementTree API [23]. La librería se encuentra bajo licencia BSD. Mientras que las librerías que extiende libxml2 y libxslt permiten su uso bajo licencia MIT.

- **Levenshnshtein**

La extensión de Levenshtein [37] para Python es una extensión desarrollada en C que permite el fácil desarrollo de operaciones como:

- Calcular la distancia de Levenshtein.
- Calcular la similitud de strings.
- Similitud de conjunto de strings.

Esta extensión posee licencia GNU.

- **TextBlob**

TextBlob [textblobWebsite](#) es una librería de Python para el procesamiento textual. Esta provee una API simple para profundizar en las tareas del procesamiento del lenguaje natural (NLP en inglés) como etiquetar partes de un discurso, análisis emocional, traducciones y más.

La librería posee una basta documentación y su licencia de uso permite el acceso, edición y uso de manera gratuita.

5.3. Características del servidor

El servidor utilizado para realizar este prototipo corresponde al servidor de prueba de Django y la máquina utilizada posee las siguientes características:

Característica	Descripción
Sistema Operativo	Ubuntu Versión 12.0(quantal) 64-bit
Procesador	Intel Core 2 Duo CPU T5870 2GHz x 2
Memoria	1.9Gb
Velocidad descarga de la red ¹	15 Mbps
Velocidad subida de la red ²	1.26 Mbps

Cuadro 5.1: Características servidor

5.4. Modelo de Datos

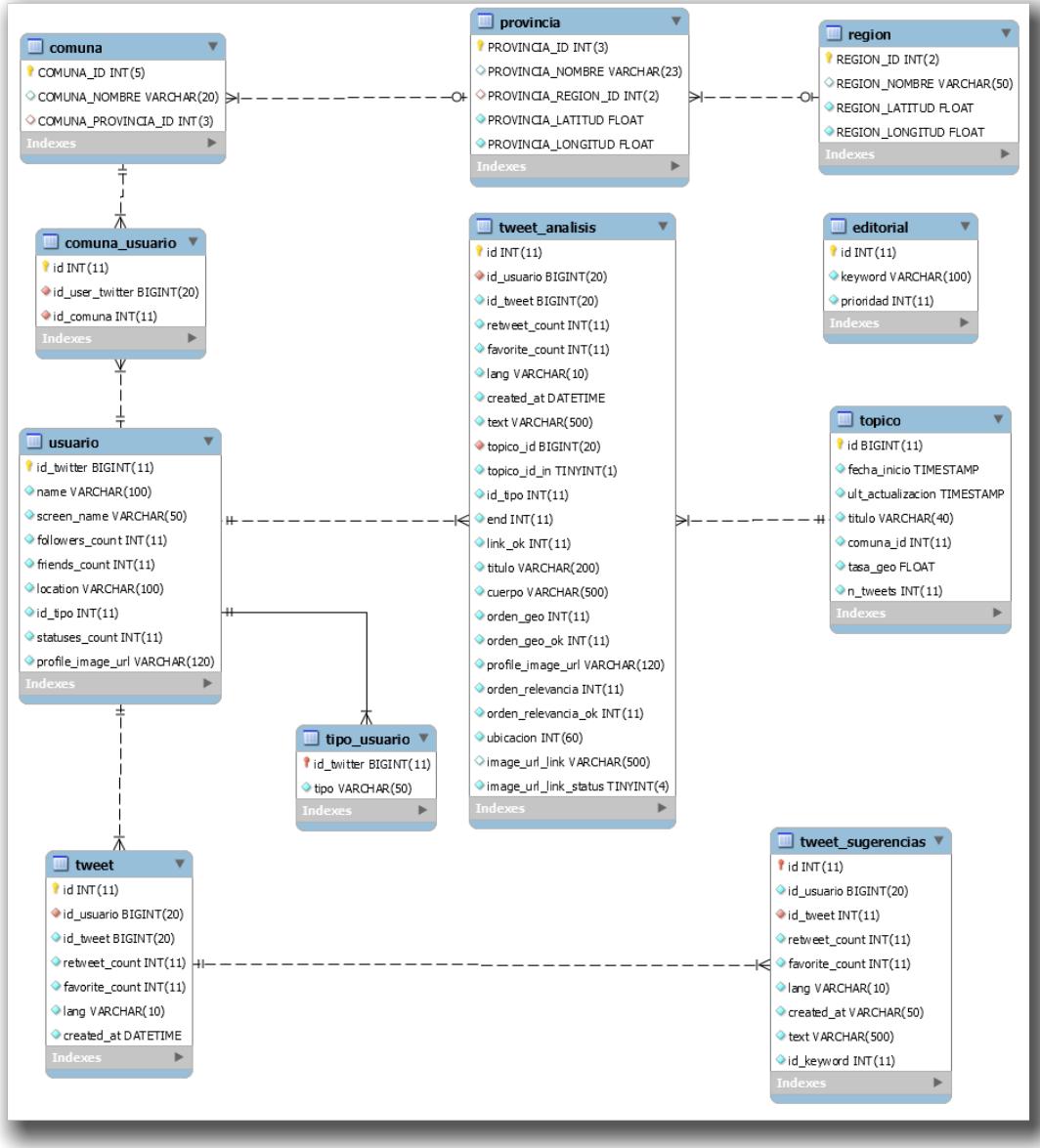


Figura 5.2: Modelo de base de datos

5.4.1. Análisis de la línea editorial del medio objetivo

Para tener una directriz respecto a qué tópicos de noticias son cubiertos por el medio de prensa objetivo, se realiza un análisis de los tweets a la fecha de dicho medio.

Este proceso se realiza mediante el conteo de la frecuencia de las palabras contenidas en

los corpus de los tweets, sin considerar las palabras vacías o stopword presentes.

Algoritmo 1 Obtención de las palabras más frecuentes del timeline de un conjunto de tweets

```

1: function GETKEYWORDMEDIO(tweets)
2:   for tweet in tweets do
3:     for tweet in tweets do
4:       for palabra in tweet do
5:         if stopwords.not_in_array(palabra) then
6:           palabras.push(palabra)
7:           frecPalabras[palabra]++
8:   return frecPalabras.ordenar()
  
```

Posición	Palabra
1	Estudiantes
2	Valparaíso
3	Universidad
4	Toma
5	Sede
6	Represión
7	Marcha
8	Chile
9	Concepción
10	Casa Central
11	Usm
12	Utfsm
13	Paro
14	Nacional
15	Carabineros
16	Movimiento
17	Pucv
18	Asamblea
19	Trabajadores
20	Secundarios

Cuadro 5.2: Tabla que muestra en orden descendente las keyword con mayor frecuencia en el análisis del timeline de Twitter de Amor TV

5.4.2. Geolocalización de usuarios

5.4.2.1. Implementación

El método implementado utilizando el concepto de la distancia de Levenshtein para determinar si el texto proporcionado por el usuario como campo de ubicación corresponde o no a una comuna existente en Chile. Cabe señalar que la distancia de Levenshtein corresponde a la cantidad de cambios necesarios en un string para transformarlo en un string objetivo.

La información relativa a las actuales regiones, provincias y regiones de acuerdo al Decreto Exento N° 817, del Ministerio del Interior, publicado en el Diario Oficial del 26 de Marzo de 2010 [14]. La posición GPS de cada una de las provincias se obtuvo mediante la ubicación proporcionada en [7].

El método implementado posee un parámetro (la máxima distancia de Levenshtein permitida) que fue probado con distintos valores para corroborar su valor más óptimo, en base a la muestra de usuarios recolectados para este trabajo y que grado de error existía en la identificación de este parámetro.

Algoritmo 2 Reconocimiento de ubicación del usuario mediante Levenshtein

```

1: function GETDISTANCIALEVEN(usuarios)
2:   for usuario in usuarios do
3:     usuario.ubicacion = limpiarPuntuacion(usuario.ubicacion)
4:     usuario.ubicacion = quitarNacionalidad(usuario.ubicacion)
5:     for comuna in comunas do
6:       dist = Levenshtein(comuna, usuario.ubicacion)
7:       if dist < MinimaDistancia then
8:         usuario.comuna = comuna

```

5.4.2.2. Resultados obtenidos

Para obtener cual es la distancia de Levenshtein con mejores resultados se tomó una muestra representativa para calcular el error. La muestra representativa consideró a 383 usuarios elegidos de manera aleatoria cuyo campo ubicación fuese distinto a vacío.

Los resultados obtenidos fueron los siguientes:

D. Levenshtein	Nº uniones	Nº uniones correctas	Error porcentual
1	157	157	0,00 %
2	172	164	2,09 %
3	207	189	4,70 %
4	223	192	8,09 %
5	235	194	10,70 %

Cuadro 5.3: Tabla comparativa para distintas distancias de Levenshtein

Podemos observar que a medida que aumentamos la distancia de Levenshtein, va aumentando la cantidad de coincidencias correctas pero también la cantidad de falsas coincidencias que ocurren.

Se consideró que un error menor al 5 % es adecuado considerando la cantidad de coincidencias correctas que aporta al conjunto. Por lo cual se utilizará la distancia de Levenshtein de 3 para realizar el match entre usuarios y comunas. Utilizando este método el número de coincidencias entre el campo ubicación y los nombre estándar de las comunas es de 114.016 del total de 650.000 usuarios (correspondiente al 17,54 % de usuarios).

Al disponer los distintos usuarios en un mapa utilizando la API de mapas de google, se obtienen las siguientes visualizaciones:

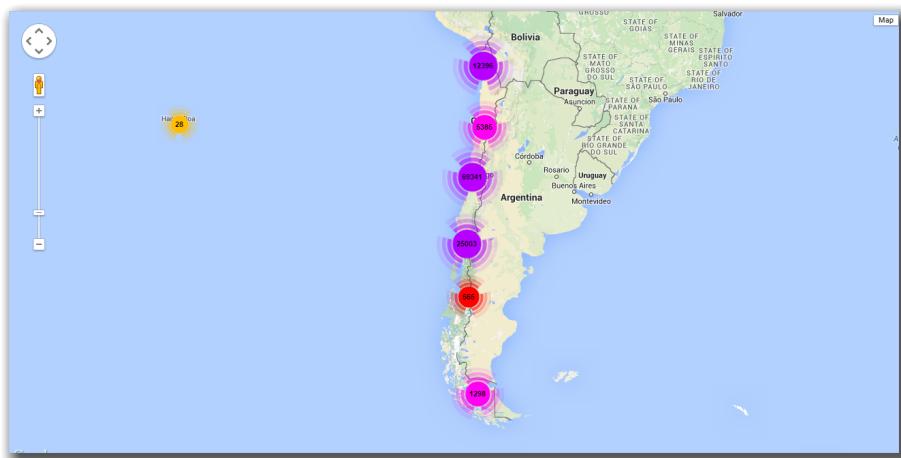


Figura 5.3: Mapa con los usuarios por ubicación geográfica

5.4.3. Captación de usuarios

El proceso de obtención del conjunto de usuarios se diseñó con la intención de obtener todos los usuarios residentes en territorio chileno. Dicho catastro no existe en ninguna fuente oficial actualizada.

El conjunto de usuarios fue creado en base a estas dos premisas:

1. Según la descripción de Java [45], los generadores de información poseen una gran base de seguidores.
2. En Twitter es habitual que los seguidores de medios de prensa al querer difundir una información le escriban un tweet a algún medio de prensa en cuestión, esperando que este realice un re-tweet, para llegar también a su base de seguidores.

El proceso de construcción de la lista de usuarios se componen de dos grandes etapas:

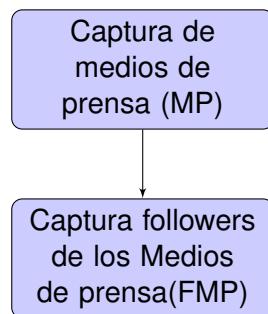


Figura 5.4: Diagrama conceptual con las etapas para la captación de usuarios.

5.4.3.1. Captura de los medios de prensa (MP)

Para generar la lista de medios(MP) se accedió a los medios de prensa registrados en las tres asociaciones más grandes de medios de comunicación de Chile:

- ANP (Asociación Nacional de Prensa) [16]: Es una asociación gremial constituida el 24 de agosto de 1951. Agrupa a los principales diarios y revistas del país. Los diarios miembros de esta asociación con cuenta en Twitter al 5 de junio del 2014 son 43.
- ANARCICH (Asociación nacional de Radios Comunitarias y Ciudadanas de Chile) [20]: Es el organismo que agrupa a 300 radios comunitarias y ciudadanas de todo el país. Las

radios que forman parte de esta asociación con cuenta de Twitter al 5 de junio del 2014 son 23.

- ARCHI (Asociación de Radiodifusores de Chile) [18]: Fundada en 1933 es la organización gremial de medios de comunicación social más antigua de Chile.

Ninguna de las asociaciones de medios de prensa considerados anteriormente (ANP, ANAR-CICH y ARCHI) cuentan con directorios públicos [17] [15] [19] que provean las cuentas oficiales de twitter de los diversos medios. Por lo cual, para recolectar las cuentas de twitter de éstos, se implementó el siguiente algoritmo ejecutado por un ser humano:

Algoritmo 3 Construcción lista de medios

```

1: function GETTWITTERACCOUNT(listaMedios)
2:   for medio in listaMedios do
3:     busqueda ← 'site:twitter.com'+medio.nombre+medio.tipo +'chile'
4:     resultGoogle ← BusquedaGoogle ( busqueda, limit = 12)
5:     for result in resultGoogle do
6:       if result.title && result.description se relacionan con medio then
7:         medio.screenName ← result.screenName
8:   return listaMedios

```

5.4.3.2. Captura followers de los medios de prensa (FMP)

Tras obtener la lista de medios de prensa, se debe proceder a obtener mediante la API de Twitter los seguidores de cada uno de los medios de prensa. El algoritmo requiere de realizar intervalos de pausas debido a las restricciones de solicitudes por hora que impone la API.

La captura de usuarios se obtuvo mediante el siguiente algoritmo:

Algoritmo 4 Captura de usuarios

```

1: function GETPOP(mediosPrensa)
2:   for medio in mediosPrensa do
3:     if GetFriendsInformation(medio, api) then
4:       Sleep(2);
5:     else
6:       Sleep(60*15);
7:       GetFriendsInformation(medio, api)
8: function GETFRIENDSINFORMATION(user, api)
9:   TwitterFriends gets api.GetFollowers(screenName=user)
10:  if TwitterFriends.length > 0 then
11:    for Friends in TwitterFriends do:
12:      SaveInBd(Friends)
13:    else
14:      Sleep(60*15);

```

Una de las dificultades presentadas en el algoritmo anterior, era que ante usuarios de más de 1,5 Millones de followers la petición por Followers se demoraba un tiempo excesivo (más de 48 horas) y retornaba error por *timeout* de la conexión. Para sortear esta dificultad fue necesario modificar la API Python Twitter directamente, agregando el retorno del cursor aún cuando se agota la conexión con la API y guardando los resultados parciales de las respuestas. El cursor de una llamada en la API es similar a un índice que permite realizar solicitudes de manera segmentada a la API³. Esta modificación fue realizada en base a las recomendaciones y comentarios disponibles en internet [56] [10].

³ Al día 26/10/2015 no se encontraba disponible la mejora en el *Github* oficial de la librería

Algoritmo 5 GetFollowers

```

1: function GETFOLLOWERS(self, screen_name=None, cursor=-1)
2:     result ← []
3:     parameters ← {}
4:     while True do:
5:         next_cursor, previous_cursor, data ← api.GetFollowersPaged(user_id, screen_name,
cursor)
6:         result += [User.NewFromJsonDict(x) for x in data['users']]
7:         if next_cursor == 0 or next_cursor == previous_cursor then
8:             break
9:         else
10:            cursor ← next_cursor
11:            sec ← self.GetSleepTime('/followers/list')
12:            time.sleep(sec)
13:    return result

```

Algoritmo 6 GetFollowers con mejora

```

1: function GETFOLLOWERS(self, screen_name=None, cursor=-1):
2:     result ← []
3:     parameters ← {}
4:     remaining ← 15
5:     ratelimited ← False
6:     while remaining > 1 do
7:         remaining ← remaining-1
8:         parameters['cursor'] ← cursor
9:         json ← api._RequestUrl(url, 'GET', data=parameters)
10:        data ← api._ParseAndCheckTwitter(json.content)
11:        if data then
12:            result += [User.NewFromJsonDict(x) for x in data['users']]
13:            if 'next_cursor' in data then:
14:                if data['next_cursor'] == 0
15:                OR data['next_cursor'] == data['previous_cursor'] then
16:                    break
17:                else:
18:                    cursor ← data['next_cursor']
19:                else:
20:                    break
21:            else
22:                ratelimited ← True
23:                break
24:    return (cursor, result, ratelimited)

```

La modificación mencionada anteriormente se puede observar en el algoritmo 6 donde se

resaltaron con otro color las líneas específicas. Principalmente incorpora un límite de llamadas a la API de 15 llamadas consecutivas, de tal manera de reducir el intervalo de respuesta y llamadas por unidad de tiempo (este parámetro fue determinado en base a la restricción de la API de Twitter). Ésta modificación permite ir guardando nuevos datos de manera continua sin necesidad de esperar hasta completar todos los followers de un usuario. En la función original las líneas 4 y 7 mantienen el ciclo de consultas por followers hasta que los cursores indican que no existen más followers por recibir (y sólo cumplida esa condición retorna datos).

El algoritmo de captura de followers se demora en promedio 2,6945531 segundos en descargar los datos de un usuario y almacenarlos en el sistema.

5.4.4. Captura de tweets

El proceso de captura de tweets se realiza obteniendo los 100 últimos tweets de cada uno de los usuarios, sin discriminación, tal como muestra el siguiente algoritmo:

Algoritmo 7 Algoritmo para la captura de tweets.

```

1: function GETTWEETS
2:   usuarios ← getUsersFromBD();
3:   for usuario in usuarios do
4:     GetUserTimeline(id_user=usuario.id);
5:     time.sleep(5);
6:   function GETUSERTIMELINE(id_user):
7:     statuses ← api.GetUserTimeline(user_id=id_user,count=100);
8:     SaveTweetInBD(statuses);
9:     if statuses.error == 34 then                                ▷ La cuenta ya no existe
10:      return 1;
11:    if statuses.error == 179 then                                ▷ La cuenta es privada
12:      return 1;
13:    if statuses.error == 88 then                                ▷ Límite de solicitudes excedidas
14:      time.sleep(5*60);
15:      return

```

El algoritmo planteado principalmente en su primera etapa realiza una búsqueda de los usuarios y sus respectivos estados referentes a si han sido recolectados sus últimos tweets (en cuyo caso se van a buscar los 100 tweets más recientes a partir del último recogido) o no (en cuyo caso se van a buscar los 100 tweets más recientes), posteriormente se almacenan en la base de datos los tweets recibidos. Es importante resaltar que este algoritmo gestiona las distintas

pausas necesarias para respetar los límites de la API Twitter.

Si la cuenta no existe o es privada se elimina del conjunto de usuarios. El algoritmo se demora en realizar una petición 0,9658139 segundos promedio.

5.4.5. Procesamiento de los tweets

El procesamiento de los tweets se define en cinco procesos:

1. Definición del tópico.
2. Obtención del conjunto de tweets relacionados al tópico.
3. Depuración de conjunto de tweets relacionados al tópico.

5.4.5.1. Definición del tópico

Un tópico son las palabras claves que definen una búsqueda temática realizada por el administrador del sistema ingresado al sistema mediante la plataforma web, cada tópico posee los siguientes atributos:

- Fecha de inicio : Se refiere a la fecha de inicio de emisión de los tweets objetivos que se requiere reunir.
- Fecha última actualización: Se refiere a la última fecha en la cual se realizó alguna modificación referente al tópico.
- Título: Se refiere a las palabras claves que definen al tópico.
- Comuna: Comuna relacionada al tópico.
- Tasa de contenidos georeferenciados: Relación entre tweets con relación geográfica y los tweets totales.
- Cantidad de tweets relacionados: Total de tweets depurados del conjunto de tweets relacionados.

5.4.5.2. Obtención del conjunto de tweets relacionados al tópico

Para obtener el conjunto de tweets relacionados al tópico se realiza una búsqueda en todos los tweets emitidos durante el periodo de interés y que contenga las palabras claves que definen al tópico. Posteriormente se eliminan los tweets que tengan una similitud 0.85 tras realizar una comparación de sus textos utilizando la librería *difflib*.

La consulta *getTweetKeyword* se demora aproximadamente entre 30 y 130 segundos.

Algoritmo 8 Obtención del conjunto de tweets relacionados al tópico

```

1: function GETTWEETKEYWORD(keywords,fecha):
2:     tweets ← getTweetsFromBD(keywords, fecha);
3:     tweets ← eliminacionTweetsRepetidos(tweets);
4:     return tweets

```

5.4.5.3. Depuración de conjunto de tweets relacionados al tópico

El proceso de depuración del conjunto de tweets obtenidos en el proceso anterior considera dos procesos distintos dependiendo del volumen de tweets involucrados.

Un gran problema en la depuración de los tweets es poder discernir si un tweet en particular se relaciona o no con la temática del tópico o si responde a una temática completamente distinta y fue relacionado con el tópico únicamente por concordancia textual. Por ejemplo, si buscamos tweets relacionados con la paralización del registro civil, con la keyword 'paro' nos gustaría filtrar todos los tweets que se refieran a un paro cardíaco o de la acción de levantarse.

Este problema fue abordado, considerando las siguientes premisas:

- El administrador del sistema posee poco tiempo.
- Es fundamental filtrar con precisión los tweets que no corresponden a la temática.

Para conjuntos de tweets de menos de 300 tweets se considera que su depuración puede ser realizada de manera manual, debido a que no es un gran conjunto de datos.

Para conjuntos de tweets de 300 tweets o más se considera que su depuración manual es muy extensa y debe ser automatizada. Para su automatización se implementó un clasificador de Bayes-Naive. El valor para determinar el límite entre la utilización entre un método y otro se realizó en base a la regla de 80-20 para el conjunto de entrenamiento.

El clasificador Bayer-Naive utilizado es una implementación de la librería Python Textblob [66], librería para el procesamiento de lenguaje natural.

Algoritmo 9 Crear Clasificador

```

1: function CREARCLASIFICADOR
2:   restoTweets, tweetsValidacion, tweetsEntrenamiento  $\leftarrow$  SepararConjuntos(Tweets);
3:   clasificador.entrenar(tweetsEntrenamiento)
4:   clasificador.validar(tweetsValidacion)
5:   for tweet in restoTweets do
6:     clasificador.clasificar(tweet);
  
```

5.4.5.4. Orden geográfico

El orden geográfico busca dotar a la herramienta de un filtro que privilegie a las fuentes cercanas al lugar indicado como origen o centro de la noticia en desmedro de aquellas que se ubican a mayor distancia geográfica.

El orden geográfico se basa en el pareo geográfico analizado previamente en la sección 5.4.2.1. Inicialmente se obtienen las comunas con menor distancia a la comuna central determinada por el tópico, la distancia se va aumentando progresivamente hasta clasificar todos los tweets de autores geoposicionados. Todos los tweets sin ubicación son desplazados al final del ranking.

Algoritmo 10 Orden Geográfico

```

function ORDENGEOGRAFICO(comunaCentral)
  d  $\leftarrow$  0
  i  $\leftarrow$  0
  tweets  $\leftarrow$  getTweetsUbicados()
  while todosAsignados(tweets) == false do
    comunas  $\leftarrow$  getComunasFromDistance(d, comunaCentral)
    for comuna in comunas do
      tweets  $\leftarrow$  getTweetsFromComuna(comuna)
      for tweet in tweets do
        tweet.ordenGeo  $\leftarrow$  i
        i  $\leftarrow$  i + 1
    d  $\leftarrow$  d + 1
  
```

5.4.5.5. Orden de relevancia

El objetivo de este *ranking*, es generar un mecanismo que ordene los contenidos en base a su relevancia, entendida como el grado de utilidad del tweet para informar al usuario respecto al evento en cuestión.

Para su diseño fue considerada la cantidad de re-tweets con que cuenta el tweet, basado en el razonamiento abordado en [77] donde se considera que la acción de re-tweet resumen en un solo indicador la importancia que le atribuyen sus lectores al contenido del tweet y lo replican, porque lo consideran relevante. Por otra parte si la intención del re-tweet es difundir (el cual es nuestro caso de búsqueda, tweets que busquen informar sobre el tópico), es bastante probable que posea una gran cantidad de re-tweet, no así los tweets conversacionales [82].

En el diseño también se consideró la fecha de emisión del tweet, implementando un sistema de tres clasificaciones basado en el ranking desarrollado en [21] con la diferencia que en se utilizaron tres grados de clasificación y no cinco.

El mecanismo de descenso de categorías implementado considera que un tweet esta *fuerza de la fecha* cuando la fecha de emisión del tweet es 4 días antes que la fecha de emisión del tweet más reciente del conjunto de tweets y considera que un tweet esta totalmente *fuerza de fecha* cuando la fecha de emisión del tweet es 10 o más días antes que la fecha de emisión del tweet más reciente del conjunto. Para los tweets *fuerza de fecha* se aplica un descenso de una categoría, mientras que para los tweets *totalmente fuera de fecha* se aplica un descenso de dos categorías.

Algoritmo 11 Orden Relevancia

```

1: function ORDENRELEVANCIA
2:   maxRT  $\leftarrow$  getMaxRT()
3:   minRT  $\leftarrow$  getMinRT()
4:   fechaMasReciente gets getFechaMasReciente()
5:   firstClass, secondClass, thirdClass gets dividirConjuntoPorRT(tweets)
6:   for tweet in firstClass do
7:     deltaFecha  $\leftarrow$  diff(fechaMasReciente, tweet.fecha)
8:     if deltaFecha  $\geq 4$  dias then
9:       tweet.clase  $\leftarrow$  tweet.clase - 1
10:    else if deltaFecha  $\geq 10$  dias then
11:      tweet.clase  $\leftarrow$  tweet.clase - 2
12:    for tweet in secondClass do
13:      deltaFecha  $\leftarrow$  diff(fechaMasReciente, tweet.fecha)
14:      if deltaFecha  $\geq 4$  dias then
15:        tweet.clase  $\leftarrow$  tweet.clase - 1

```

5.4.5.6. Panel de enlaces

Esta vista fue desarrollada con la intención de crear una sección de *bibliografía multimedia* de un tópico. Mediante un algoritmo que recoge todos los enlaces compartidos mediante los tweets del tópico y los presenta a modo de grilla.

Esta vista permite en mismo espacio, acceder a enlaces externos que permiten extender la información del tópico aportada por los distintos tweets, así como acceder a enlaces de distintas posturas permitiendo entre otras cosas contraponer la forma en que presentan la información, profundidad, etc.

Los enlaces son ordenados en base a su tweet de origen, del más reciente al menos reciente y luego del que posee más re-tweets al que posee menos.

Algoritmo 12 Obtención de enlaces externos contenidos en los tweets

```

function GETURLSBYTWEETS(tweets) tweets  $\leftarrow$  ordenarRelevancia(tweets);
  for tweet in tweets do
    if hasURL(tweet) then
      urls.append(tweet.url)
  for url in urls do
    url.titulo, url.link, url.imagen  $\leftarrow$  getInformationURL(url)

```

Debido a la gran versatilidad del origen que poseen los distintos enlaces compartidos, existen dificultades en su recolección por distintos motivos como: URL incorrectas o inexistentes,

textos que no es posible manipular, etc.

Para capturar los datos del enlace, se utilizó la librería lxml 5.2 de Python, navegando por la estructura del documento HTML se extrae el título, en enlace completo y una imagen representativa. Las condiciones mínimas que se aplica a cada uno de estos datos, es que el título tenga más de 10 caracteres, que la imagen posea el metatag *image* de OpenGraph [24] y que el enlace no retorne error 404 (página no encontrada).

5.4.5.7. ON/OFF Medios de prensa

El botón *ON/OFF Medios de prensa* permite ocultar o mostrar todos los tweets de la lista que hayan sido emitido por una cuenta registrada en el sistema como medio de prensa. Esta funcionalidad fue desarrollada con la intención de contar con la opción voluntaria de visualizar o no, los tweets de las cadenas de prensas, para privilegiar la lectura de tweets generados por personas o entidades sociales.

Capítulo 6

Evaluación y discusión

6.1. Vistas del prototipo

- Vista nueva búsqueda tópico: Esta vista presenta una barra de búsqueda para ingresar las palabras claves del tópico, la fecha inicial del tópico y un select box para la ubicación del suceso.

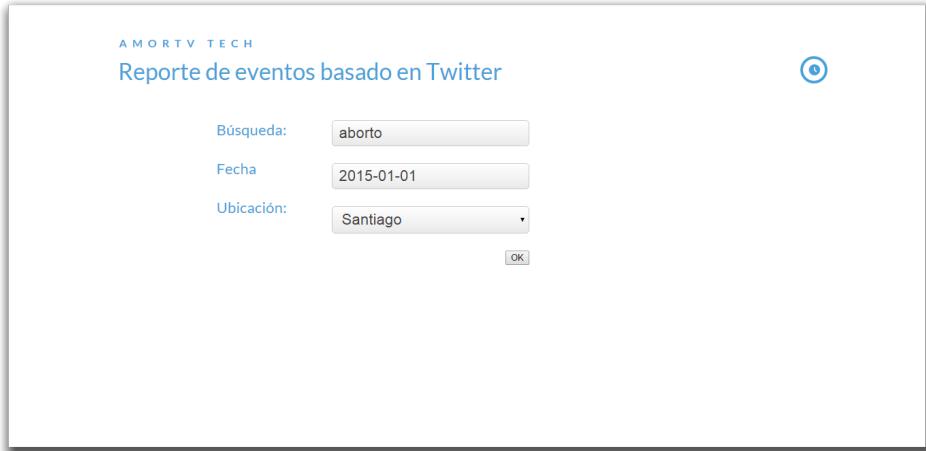


Figura 6.1: Vista nueva búsqueda tópico

- Vista resultados tweets para nuevo tópico: Esta vista incluye la cantidad de tweets relacionados al tópico que contabilizó el algoritmo.



Figura 6.2: Vista resultados tweets para nuevo tópico

- Vista tópicos: Esta vista presenta la lista de los tópicos que han sido buscados previamente, en cada tópico se indica distintas características del tópico: cantidad de tweets, fecha de inicio y ubicación.

The screenshot shows a user interface titled "Reporte de eventos basado en Twitter". At the top right, there is a circular icon with a person symbol. Below it, the text "EVENTOS REPORTEADOS" is displayed. The main content area contains a grid of ten blue boxes, each representing a topic:

[1] Portuario @[sin_ubicacion] #Tweets:710 Jan. 1, 2014, 3 a.m.	[2] Sol Lluvia @[sin_ubicacion] #Tweets:33 July 1, 2014, 4 a.m.
[3] Confech @[sin_ubicacion] #Tweets:191 July 30, 2014, 11:58 p.m.	[4] Camila Vallejos @[sin_ubicacion] #Tweets:52 Aug. 17, 2014, 4:52 a.m.
[5] Educacion @[sin_ubicacion] #Tweets:141 Aug. 17, 2014, 9:44 p.m.	[6] Paro @[sin_ubicacion] #Tweets:711 Oct. 28, 2014, 12:17 a.m.
[7] Camila Vallejos @[sin_ubicacion] #Tweets:77 Nov. 15, 2014, 11:12 p.m.	[8] Paro Docente @[sin_ubicacion] #Tweets:342 Dec. 14, 2014, 2:14 p.m.
[9] Machismo	[10] Feminism

Figura 6.3: Vista de tópicos

- Vista orden geográfica: Esta vista presenta un listado de los tweets relacionados en orden geográfico, en la parte superior de la lista están los tweets emitidos en la ubicación del tópico mientras que en la parte inferior de la lista están los tweets emitidos en la comuna más distante físicamente de la ubicación del tópico.

The screenshot shows a user interface titled "Reporte de eventos basado en Twitter". At the top right, there is a circular icon with a person symbol. Below it, the text "ABORTO [ORDEN GEGRÁFICO] PRENSA ON" is displayed. The main content area shows a vertical list of tweets with their times and dates:

2014-07-24 18:03	@uchileradioONU sugiere a Chile modificar la Ley Antiterrorista y despenalizar el aborto en caso de una... Iquique
2014-11-04 10:12	@imfmonEs menor, fue violada, quedó embarazada, no puede pagar un aborto, es obligada a parir un feto inviable. ¿Cuántas formas de v... Iquique
2015-02-03 12:39	@ReporteChileEstablecimientos médicos del Opus Dei se suman a posición de rector UC sobre aborto http://t.co/CVoRWCWfO Iquique
2015-02-03 09:25	@AhNoticiasMegaGinecólogo UC y proyecto de abortoSe va a legalizar matar seres humanos indefensos Iquique
2015-01-31 22:26	@pvzamora#LaVida #NoAlAborto Porque todo Chile defiende a sus hijos, dignidad a las mujeres embarazadas y los bebés por nacer! Iquique
2014-06-30 12:17	@elmostradorDiputadas, alcaldesas y dirigentes sociales formarán coordinadora pro despenalización del #Aborto. http://t.co/tGDMsrc5fw Antofagasta

Figura 6.4: Vista de orden geográfico

- Vista orden por relevancia: Esta vista presenta un listado de los tweets relacionados en orden descendiente por relevancia.



Figura 6.5: Vista de orden por relevancia

- Vista orden temporal: Esta vista presenta un listado de los tweets relacionados en orden descendiente de fecha de emisión.



Figura 6.6: Vista de orden temporal

- Vista links: Esta vista presenta un listado de los enlaces externos contenidos en los distintos tweets relacionados con el tópico, cada una de las casillas presenta una imagen previa del contenido y el título del contenido del enlace. Las casillas blancas mostradas en la imagen corresponden a que la imagen previa no pudo ser obtenida.

The screenshot displays a grid of news items from Twitter. The top row features a user profile picture, a Chilean flag, and a close-up of a baby's hand. The bottom row shows a portrait of a man and a red logo with white letters. The interface includes a header 'AMORTV TECH' and 'Reporte de eventos basado en Twitter', along with various sharing icons.

Figura 6.7: Vista de links

- Botón ON/OFF Prensa: Este botón oculta o muestra los tweets emitidos por las cuentas identificadas como medio de prensa.

The screenshot shows a chronological timeline of tweets about abortion, filtered for 'PRENSA OFF'. The tweets are timestamped and include user profiles and their tweets. The interface includes a header 'AMORTV TECH' and 'Reporte de eventos basado en Twitter', along with various sharing icons.

Figura 6.8: Vista de tópicos

6.2. Caracterización de la población de datos capturados

En la siguiente sección se presenta una caracterización general de los datos capturados hasta el cierre de esta memoria con intenciones de aportar a los datos estadísticos relativos a Twitter en territorio Chileno.

Tipo de dato	Nº
Usuarios	650.000
Tweets	17.300.000

Cuadro 6.1: Cantidad de datos capturados

Características de uso de Twitter

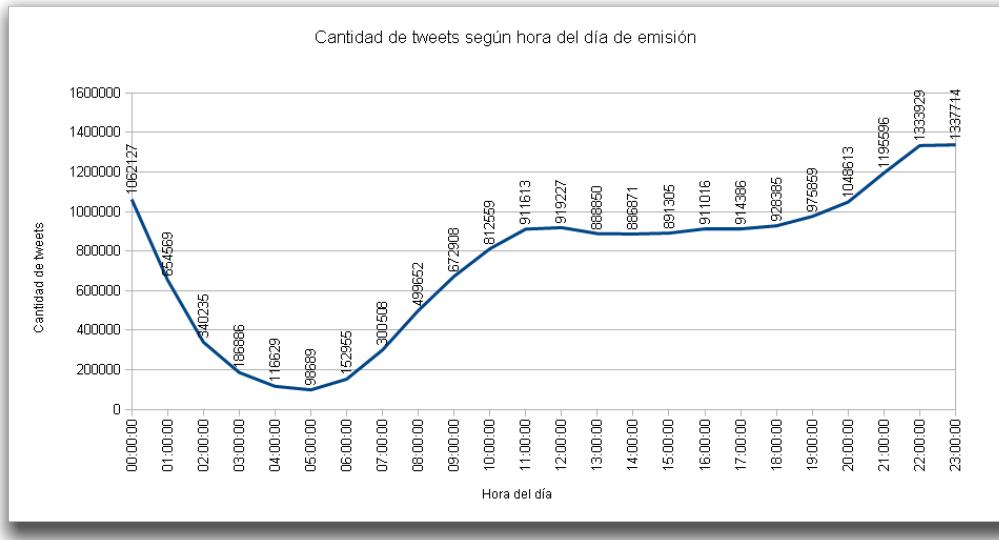


Figura 6.9: Distribución de tweets por hora de emisión

En la figura 6.9 se grafica la distribución horaria de la emisión de los tweets captados se puede observar que las horas donde se emiten mayor cantidad de tweets comprende el periodo del día que va desde las 20:00hrs. hasta las 23:00hrs. mientras que el de menor emisión de tweets comprenden el periodo de horas desde las 3:00hrs. hasta las 6:00hrs.

Se observa además que existe un periodo relativamente constante en emisión de tweets que se desarrolla desde las 11:00hrs. hasta las 20:00hrs.

La figura 6.10 representa la actividad porcentual por horas separado por región, donde se observa que existe un comportamiento similar al descrito anteriormente.

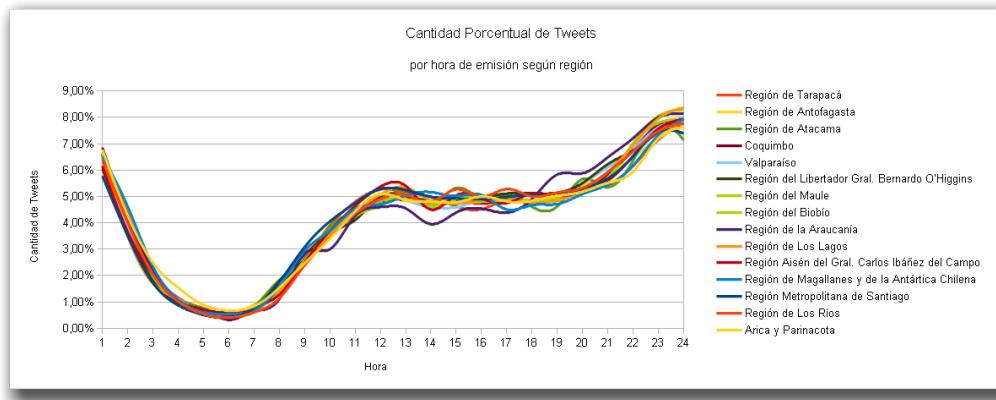


Figura 6.10: Distribución porcentual de tweets por hora de emisión según región

Cantidad de re-tweets por usuario

Medida	Nº
Moda	0
Rango	3.424.962
Desviación Estándar	26.900,05
Promedio	711,31

Cuadro 6.2: Datos cuantitativos respecto a los RT

La tabla 6.2 nos permite observar que en cuanto al parámetro de RT el conjunto de tweets existe una variabilidad increíble respecto al promedio, lo que sugiere que existen tweets particulares con cantidades extremas de re-tweets. Tras analizar con más detalles los tweets con mayores cantidades de re-tweets se identifica una hecho insólito incluso para las macrocifras de Twitter: el tweet con más re-tweets corresponde a la famosa *selfie* tomada por Ellen DeGeneres en los Oscar 2014 en conjunto a varias estrellas del cine que alcanzó el record en re-tweets con la suma de 2,5 millones de re-tweets.



Figura 6.11: Tweet que generó nuevo record del mensaje más re-tweeteado.

Cantidad de favoritos por usuario

Medida	Nº
Moda	0
Rango	18423
Desviación Estándar	16,28
Promedio	0,2602

Cuadro 6.3: Datos cuantitativos respecto a los FAV

De manera similar al caso de los re-tweets, la cantidad de favoritos presenta una desviación estándar más grande que el promedio captado, con la diferencia que el rango es menor. En comparativa con la tabla anterior es posible apreciar una diferencia de al menos tres ordenes de magnitud entre las medias de ambos cuantificadores.

Cantidad de tweets por usuario

Respecto a la cantidad de tweets se observa que el promedio son 627 tweets por cuenta de usuario como se observa en la siguiente tabla:

Medida	Nº
Moda	0
Rango	718.249
Desviación Estándar	4.700,45
Promedio	627,5728

Cuadro 6.4: Cantidad de tweets por usuario

Cantidad de Seguidores por usuario

Medida	Nº
Moda	1
Rango	5.372.178
Desviación Estándar	16.504,96
Promedio	427,64

Cuadro 6.5: Cantidad de seguidores por usuario

Cantidad de Amigos por usuario

Medida	Nº
Moda	40
Rango	1.393.022
Desviación Estándar	5.283,16
Promedio	364,50

Cuadro 6.6: Cantidad de amigos por usuario

Tras comparar las tablas 6.5 y la tabla 6.6 se observa que existe una diferencia muy ajustada en los promedios obtenidos, pero que el valor máximo es superior en el caso de la cantidad de seguidores que en la cantidad de amigos de los usuarios.

Utilizando las clasificaciones para usuarios realizadas en [77] en Chile su distribución es la siguiente:

Nombre Categoría	Nº
Élite Global	784
Élite Local	1.611
Usuario Corriente	624.564

Cuadro 6.7: Cantidad de usuarios según clasificación realizada en [77]

Cómo dato adicional a la tabla 6.7 existen 37.808 que teniendo menos de 1000 followers no cumplen con el criterio de tener menos de 1000 amigos.

Posicionamiento geográfico de los usuarios

Respecto al posicionamiento de los usuarios con el método revisado en 5.4.2 se obtuvo la siguiente distribución por regiones:

Región	Usuarios Twitter		Población real	
	Usuarios (M)	Porcentaje	Personas (M)	Porcentaje
Región de Arica y Parinacota	1,914	1,68	185,0	1,1
Región de Tarapacá	4,477	3,93	314,5	1,8
Región de Antofagasta	6,005	5,27	575,3	3,4
Región de Atacama	0,974	0,85	280,5	1,6
Región de Coquimbo	4,411	3,87	718,7	4,2
Región de Valparaíso	7,111	6,24	1.759,2	10,3
Región de O'Higgins	4,756	4,17	883,4	5,2
Región del Maule	6,183	5,42	1.007,8	5,9
Región del Biobío	16,862	14,79	2.036,4	11,9
Región de la Araucanía	0,825	0,72	970,4	5,7
Región de Los Ríos	2,245	1,97	379,7	2,2
Región de Los Lagos	4,891	4,29	836,3	4,9
Región de Aisén	0,565	0,5	104,8	0,6
Región Magallanes y la Antártica	1,298	1,14	158,7	0,9
Región Metropolitana de Santiago	51,499	45,17	6.883,6	40,3
Total	114.016	100	17.094,3	100

Cuadro 6.8: Distribución de los usuarios por regiones

En esta tabla es posible observar la concentración de usuarios existente en torno a la región metropolitana con el 45 % de los usuarios totales del país, 5 puntos porcentuales menos con respecto a la población real de Chile. Lo anterior no refleja otra cosa que la concentración del país también se ve reflejada por la cantidad de usuarios de Twitter en Chile.

6.3. Resultados

Este capítulo del presente trabajo está enfocado en obtener métricas relativas a los objetivos planteados en el Capítulo 2, es por tanto que la obtención de resultados abarcará los siguientes

aspectos:

- Referencias a noticias
- Medición de componentes de opinión

6.3.1. Referencias a noticias

Tal como se definió anteriormente, se entiende por noticia Una noticia es la comunicación de información seleccionada sobre un evento actual que posteriormente es presentado a través de cualquier medio de comunicación existente :[63] Esta sección busca cuantificar el grado de noticias relacionadas por el algoritmo realizado para un tópico.

Consideramos como características relevantes de una noticia los siguientes aspectos:

- Que surja respecto a un hecho concreto ocurrido.
- Que sea presentada por algún medio de prensa.
- Que se encuentre en un margen temporal cercano a la fecha ocurrida del suceso gatillante de la noticia.

6.3.1.1. Medio de prensa Modelo

El medio de prensa modelo es considerado es una medio de prensa de referencia en los medios digitales de prensa en el territorio Chileno, este medio modelo es considerado a fin de obtener resultados comparativos.

Para su selección se utilizaron los siguientes criterios:

- Foco noticioso.
- Segmento socio-económico de su público objetivo.
- Cantidad de seguidores en Twitter.
- Actividad en Twitter.

Basándonos en el estudio [57] realizado por la Escuela de Periodismo de la Universidad Alberto Hurtado en el sitio web *puro periodismo* que describe: “Corresponden a sitios de noticias generalistas, de publicación constante, ubicados en los primeros lugares del ranking Alexa y considerando seguidores de Twitter y Fans en Facebook.”¹

Nombre	Cuenta Twitter	Menciones	Seguidores	Siguiendo	Tweets	Creación
24 Horas	@24horasTVN	3733	1.669.772	162	246.403	15/11/2009
CNN Chile	@CNNChile	1599	1.424.580	875	124.711	19/12/2008
BioBioChile	@biobio	6311	1.392.781	15.653	431.895	3/05/2008
Cooperativa	@cooperativa	3765	1.348.441	348.406	448.498	23/07/2007
La Tercera	@latercera	2905	904.200	245.123	265.246	2/04/2007

Cuadro 6.9: Cuadro descriptivo del estado de las cuentas de Twitter de los principales medios de prensa de Chile, elaboración *puro periodismo*, actualizada 26 de junio 2014.

De estos medios se seleccionaron los tres medios más populares en Twitter que provengan de métodos de difusión distintos: Televisión, Radio y Prensa Escrito. Por lo cual los medios de prensa seleccionados son: 24 Horas (Televisión mediante la señal de TVN), Bio-Bio Chile (Radio) y La Tercera (Prensa escrita).

6.3.1.2. Referencia a noticias

La referencia a noticias se aborda mediante la siguiente pregunta *¿ Cuántos tweets del conjunto seleccionado se refieren directamente a un hecho noticioso difundido por el medio modelo de prensa convencional?*

La pregunta anterior aborda la característica de una noticia, si es que es replicado por algún medio de prensa y si su surgimiento se refiere a un hecho concreto.

Se entiende por *referencia directa* cuando el texto de un tweet se refiere de manera específica y explícita a alguna información o temática y por *hecho noticioso difundido por la prensa convencional* (acorde a la definición de noticia realizada al inicio de este capítulo) se entiende específicamente como el hecho noticioso ocurrido en territorio del estado Chileno que haya generado una nota del medio modelo considerado para este estudio.

El procedimiento para definir si el texto del tweet se refiere directamente a un hecho noticioso es el siguiente:

¹El ranking de Alexa es un ranking de tráfico con potentes herramientas comparativas y de monetización para sitios web realizado por www.alexa.com (filial de Amazon)

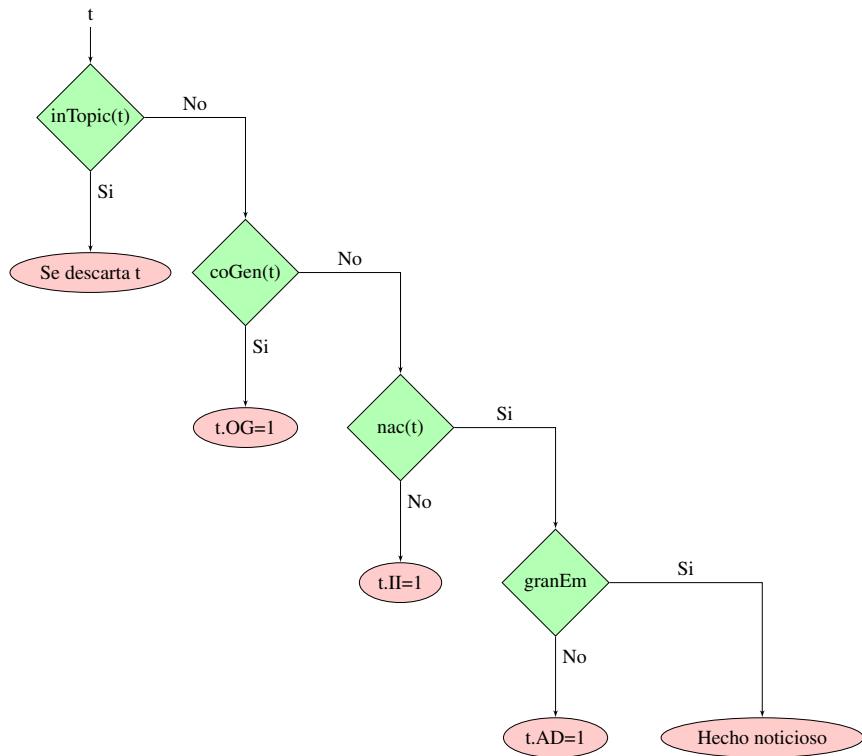


Figura 6.12: Procedimiento para determinar si un tweet se refiere a un hecho noticioso

Sub-procesos

inTopic : ¿El tweet se refiere al tópico en cuestión?

coGen : ¿El tweet se refiere al tópico en general o se refiere a un hecho en particular?

nac : ¿El hecho aludido por el tweet es nacional o internacional?

granEm: ¿El hecho posee gran envergadura?

Algoritmo 13 Procedimiento para determinar si un tweet se refiere o no, a un hecho noticioso

```

1: function REFHECHONOTICIOSO(tweet):
2:   if inTopic(tweet) then
3:     return false;
4:   else
5:     if coGen(tweet) then
6:       return tweet.OG=1;
7:     else
8:       if nac(tweet) then
9:         if granEm(tweet) then
10:          return tweet.AD = 1;
11:        else
12:          return true;
13:      else
14:        return tweet.II=1;

```

Por su parte el procedimiento para verificar si el hecho noticioso ha generado una nota del medio modelo considerado para este estudio es el siguiente:

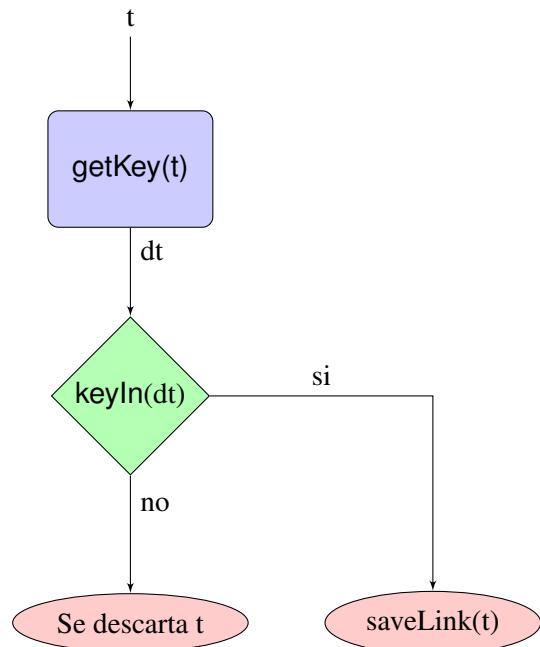


Figura 6.13: Procedimiento para verificar si el hecho noticioso haya generado una nota del medio modelo

Sub-procesos

getKey : Identificar las dos keywords mas relevantes del tweet
keyIn : Busca en el arreglo de titulares de la tercera si se encuentran las keywords
saveLink(t): Almacena el enlace de la noticia relacionada

Algoritmo 14 Pseudo-código para verificar si el hecho noticioso haya generado una nota del medio modelo.

```

1: function GENERATEENLACE(tweet):
2:   dt = getKey(tweet)
3:   if url = keyIn(dt) then
4:     return url;
5:   else
6:     return null;
  
```

Ambos procedimientos anteriores son ejecutados por un ser humano respecto debido a la dificultad existente con las fuentes de verificación. Las únicas fuentes que poseen en línea los medios modelo online de notas históricas de acceso público, son las sus respectivas secciones de búsqueda en sus sitios web respectivos [41] [9] [11].

Para poder analizar la información provenientes de estas fuentes se realizaron algoritmos de *scraping* directamente sobre los sitios web que transformaban los resultados en tablas para la comprobación de resultados.

6.3.2. Casos de prueba: El aborto

Este caso de prueba analiza el tópico noticioso referente al aborto, temática contingente que tomó la agenda nacional luego del anuncio presidencial de Bachelet respecto a la reforma para su legalización.

El tópico comprende el periodo entre 1 de junio del 2014 y el 18 de febrero de 2015. Que comprende desde la primera etapa del anuncio del proyecto de ley de despenalización del aborto hasta el inicio de análisis de este tópico.

Durante este proceso ocurrieron los siguientes hitos noticiosos contenidos en los tweets en todos los medios de prensa modelo (explicados en 6.3.1):

Fecha	Hito
2 de junio 2014	Comisiones de salud de las cámaras de diputados y senadores inician debate sobre proyectos de despenalización de aborto
26 de junio 2014	Anuncio del proyecto de despenalización del aborto
24 de julio 2014	ONU pide a Chile incluir violación como causa para hacerlo de forma legal
3 de noviembre 2014	Caso de joven de 13 años violada y embarazada
30 de diciembre 2014	Declaraciones de ministra de salud Henia Molina en la Segunda sobre los abortos en <i>clínicas cuicas</i>
30 de diciembre 2014	Renuncia de ministra Henia Molina
17 de enero 2015	DC afirma que sus votos no están asegurados para apoyar la despenalización del aborto
1 de febrero 2015	Declaraciones Rector PUC respecto a trabajadores que quieran realizar abortos en la Red UC
5 de febrero 2015	Red de clínicas privadas declara que no realizará abortos en sus recintos
6 de febrero 2015	Críticas a dichos de Lorenzini sobre los motivos, que a su juicio, causarían una violación
9 de febrero 2015	Cadem: 71 % aprueba proyecto de aborto enviado por el gobierno

Cuadro 6.10: Hitos noticiosos presentes en todos los medios de prensa modelos.

6.3.2.1. Procesamiento del tópico

Durante el proceso de entrenamiento del clasificador se consideraron los siguientes criterios para realizar la selección:

- Utilización del verbo abortar en un contexto no relacionado al tratamiento médico en cuestión.

- Insultos discriminatorios y agresiones racistas.
- Comentarios ambiguos que no permiten relacionar con el concepto aborto en cuestión.

Las cantidades según clasificación manual realizados en el proceso son los siguientes:

Clasificación	Aceptados	Descartados	Total
Entrenamiento	181	19	200
Validación	93	6	99

Cuadro 6.11: Clasificación referencia hecho noticioso

Con esta información el clasificador del total de 1382 tweets aceptó 1357 tweets.

6.3.2.2. Análisis de muestra representativa

Para analizar el contenido de los tweets seleccionados fue recogida una muestra representativa de tweets calculada considerando las siguientes variables:

- Tamaño muestra $N = 1357$
- Error estándar 15 %
- Porcentaje estimado de la muestra $P = 0,9$

El tamaño de la muestra representativa de tweets para este tópico es de 309 tweets los cuales fueron seleccionados de manera aleatoria mediante el ordenamiento pseudoaleatorio entregado por la función *RAND* de Mysql.

6.3.2.3. Análisis de contenido

Para realizar el análisis de contenido noticioso, se aplicó a la muestra el algoritmo explicado en 6.12. Con el cual se obtuvo la siguiente clasificación de los tweets de la muestra:

En la tabla 6.13 se observa los efectivos resultados del filtro que determina si el tweet corresponde o no al tópico, donde sólo el 0,5 % de los tweets no están relacionados al tópico.

Clasificación	Total	Porcentual
Tweets que hacen referencia	108	34,95 %
Tweets que no hacen referencia	200	64,72 %
	309	100 %

Cuadro 6.12: Clasificación según referencia hecho noticioso

Clasificación	Total	Porcentual
Opinión general (OG)	147	73,13 %
Insumo Internacional (II)	16	7,96 %
Aporte a la discusión (AD)	37	18,41 %
No se refiere al tópico	1	0,50 %
	201	100 %

Cuadro 6.13: Clasificación de los tweets que no hacen referencia a un hecho noticioso

La efectividad de este filtro es fundamental para las etapas de clasificación posteriores como la recolección de enlaces y análisis geográficos.

El alto porcentaje de tweets del tópico que no se refieren a un hecho noticioso comprenden las distintas formas que frecuentan los usuarios para participar del debate público en Twitter: 73,3 % de estos tweets corresponden a una opinión personal, el 18,41 % entrega algún aporte a la discusión y el 7,96 % corresponden a información o insumos internacionales.

Estos resultados son alentadores en cuanto a la completitud y diversidad de esta información pues comprenden opiniones personales, enlaces externos, contexto internacional y referencias directas a hechos noticiosos.

Clasificación	La tercera	24 Horas	Bio-Bio			
Tweets con referencia a h.n.	49	15,86 %	50	16,18 %	45	14,56 %
Tweets sin referencia a h.n.	59	19,09 %	58	18,77 %	63	20,39 %
Total	108	34,95 %	108	34,95 %	108	34,95 %

Cuadro 6.14: Cobertura de hechos noticiosos (h.n.) contenidos en los tweets en los medios de prensa (Las cantidades porcentuales son respecto al total de la muestra)

Al observar la tabla anterior 6.14 se observa que cerca del 50 % de los hechos noticiosos a los que se refieren los tweets no son cubiertos por alguno de los medios modelo, estos resultados expresan que es posible informarse de sucesos que no logran captar el interés del medio modelo para generar una nota de prensa logrando superar sus procesos de gatekeeping. De esto último podemos concluir que informarse de Twitter, permite informar de eventos que no son

difundidos por los medios de prensa modelo.

El conjunto de tweets recogidos presenta 613 enlaces externos.

6.3.2.4. Análisis temporal

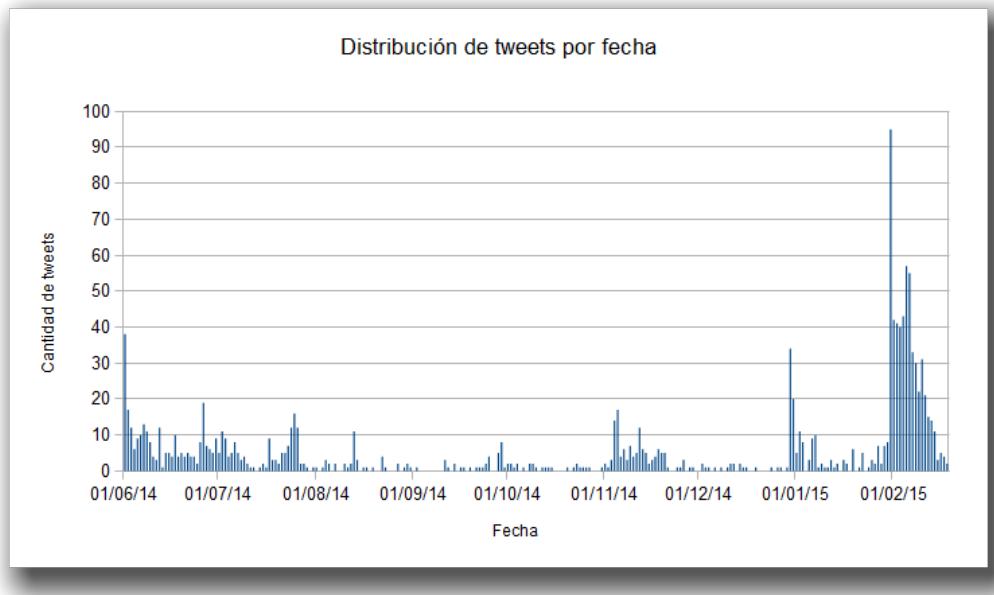


Figura 6.14: Distribución de cantidad de tweets por día

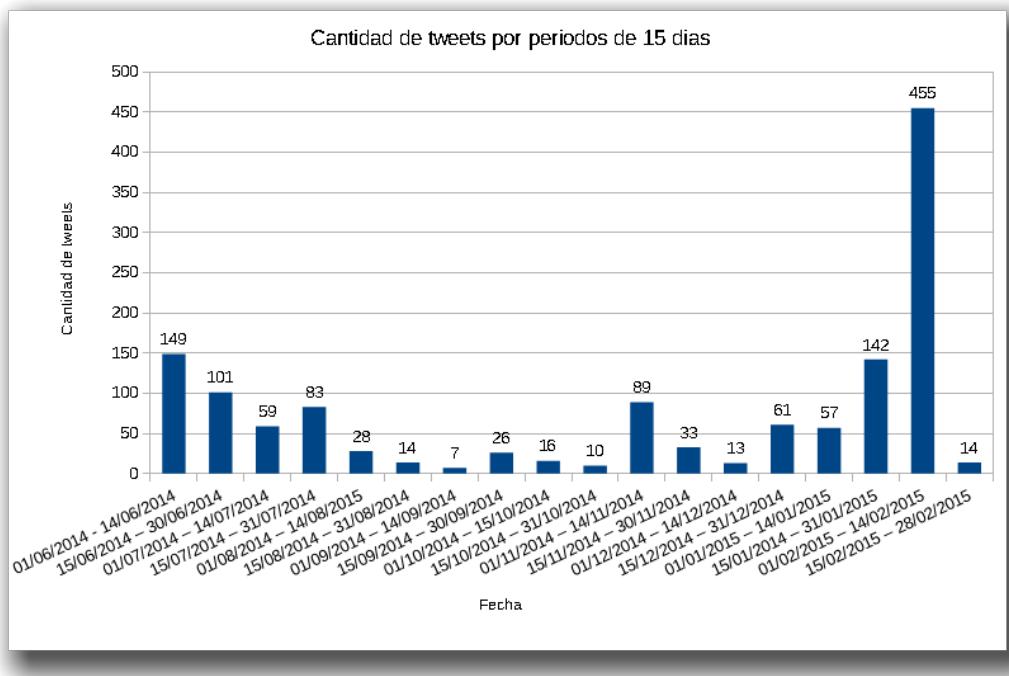


Figura 6.15: Distribución de cantidad de tweets por periodos de quince días

En el gráfico 6.14 es posible analizar la concentración de tweets por fecha, que en complemento del gráfico 6.15 Nos permiten observar que existe una gran concentración de tweets en la primera quincena de febrero 2015 que se puede relacionar con la concentración de cuatro hitos noticiosos ocurridos desde el 1 de febrero hasta el 9 de febrero (ver tabla 6.10). El segundo periodo con más tweets corresponde a la primera quincena donde ocurrió un hito noticioso el 2 de junio de 2014. El tercer periodo con más tweets contiene un hito noticioso.(y dos hitos cercanos del periodo anterior, puesto que ocurrieron dos días antes del día previo del término del periodo).

Es posible observar entonces que el debate se activa (y aumenta la cantidad de producción de tweets relacionados) en la medida que ocurren hechos noticiosos relevantes a nivel nacional.

6.3.2.5. Análisis de re-tweet

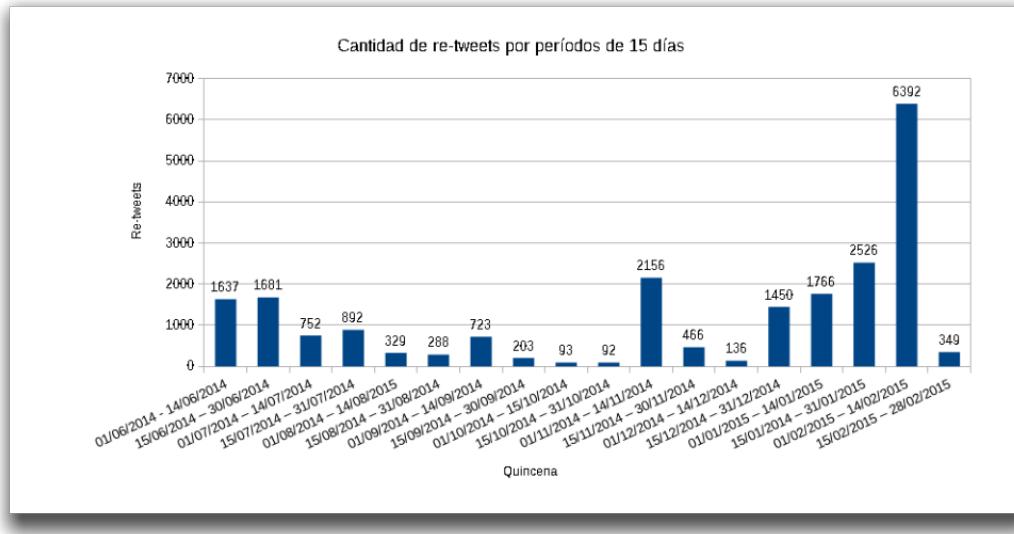


Figura 6.16: Cantidad de re-tweets por quincena

En el gráfico 6.14 es posible analizar la concentración de tweets por periodo de tiempo, separados por quincena, el periodo de mayor concentración de tweets ocurre en la quincena 17 de manera similar a la mayor concentración de tweets analizados anteriormente, la segunda concentración ocurre en la segunda quincena de enero 2015 donde ocurre un hito noticioso, mientras que la tercera mayor concentración se ubica en la primera quincena del mes de noviembre 2014, donde se ubica un hito noticioso.

Según el razonamiento desarrollado para el *ranking* de relevancia en 5.4.5.5. Los tweets con más RT son los que presentan información útil o relevante para quienes leyeron esos mensajes y realizaron la acción de re-tweet.

Nº RT	Autor	Texto de contenido
1086	@bairdCampbell	RT @fromerod: Los que prohibían el aborto, abortaban en Londres. Hoy, quienes prohíben la libertad de expresión, se expresan en París.
664	@RosarioAlcaldeG	RT @joseantoniookast: Senador Lagos Weber tildó de “rasca” volante de la UDI sobre el aborto. Viendo esta foto sólo decir: Mira quien habla !
629	@MAURYZS	RT @biobio: ONU recomienda a Chile permitir aborto a menores de 18 años por “salud fisiológica y mental”
588	@Beriiitha	RT @jmfmoran: Es menor, fue violada, quedó embarazada, no puede pagar un aborto, es obligada a parir un feto inviable. ¿Cuántas formas de v...
360	@raul_torres79	RT @link_anarquista: Chile: Niña de 13 años embarazada por violación es castigada por el Estado sin derecho a aborto
332	@csuarezespinoza	RT @DerechaTuitera: Levantemos la mano todos los que creemos que el aborto provoca daños sicológicos irreparables! (acá un ejemplo) http
253	@RuubiaNaturaal	RT @lasultimas: Van Rysselberghe: “El aborto terapéutico es un control de calidad a la raza humana”. Envía “weona loca.” y recibe chi
252	@urpiestrada	RT @PorAbortoLegal: Quién decide sobre un #Aborto? explicación sencilla http://t.co/cP4ApMZw33 #provida
243	@irutherf	RT @KennethOficial: ¡Que ironía! Los que están a favor del aborto... nacieron.
242	@zarasenda	RT @SomosDocumental: Cine militante y documental contra la reforma de la Ley del aborto. http://t.co/NDfoNF4ABM #TendráLaLibertad

Cuadro 6.15: 10 Tweets mas re-twiteados del tópico

Si analizamos los tweets en 6.15 es posible observar que 7 de los 10 cuentan con una imagen adjunta al tweet. De éstos 4 se refieren a un hecho noticioso, 4 se refieren a opiniones personales, 1 aporte a la discusión y 1 correspondiente a información internacional sobre el tema.

Número de tweets emitidos por un mismo usuario	Número de usuarios distintos
1	602
2	130
3	41
4	27
5	4
6	9
7	4
8	3
9	1
10	2
11	3
12	1
15	1
16	2
17	1

Cuadro 6.16: Cantidad de usuarios distintos por número de tweets emitidos

6.3.2.6. Análisis geográfico

La distribución de usuarios que realizaron tweets en el tópicos, con una ubicación identificada fueron 223 de los 831 autores distintos (26,8 %). De los cuales 20 se ubican en Antofagasta, 153 en Santiago,

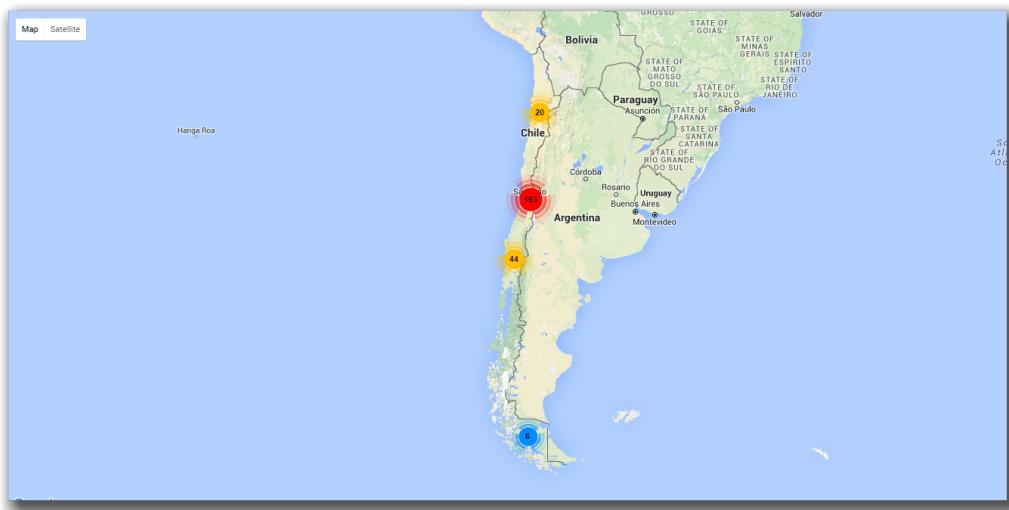


Figura 6.17: Distribución geográfica de los usuarios

Capítulo 7

Conclusiones

7.1. Conclusiones

El presente trabajo abarca aspectos teóricos y prácticos sobre la problemática de informarse sobre un evento noticioso y profundiza en el desarrollo de una algoritmo computacional que recolecta información desde Twitter y la presenta en una interfaz web, con varias alternativas de presentación de la información.

La interfaz del prototipo posee una barra superior que de manera simple permite acceder a las distintas vistas de ordenamiento de los tweets y enlaces. Los tweets se presentan de manera descendente en una línea de tiempo con enlaces directos tanto al perfil del autor como a los distintos tweets facilitando el acceso directo a la fuente original, entregando una experiencia simple e intuitiva.

En la sección 3 se analizan los distintos trabajos realizados sobre Twitter que tuvieran relación con aspectos abarcados en este trabajo tanto en estudios teóricos existentes como herramientas de fines similares. De forma general, la cantidad crecientes de estudios sobre Twitter permite verificar el creciente interés de la comunidad científica sobre esta red social.

Muchos de los trabajos revisados abordan de distinta manera la categorización de la información para dar solución a la problemática del valor de la información frente a los grandes volúmenes de Twitter. Para la clasificación de usuarios las estrategias variaban en la consideración de distintas características relacionadas a los tweets o a los usuarios en la plataforma, las más efectivas fueron aplicadas completa o parcialmente para el diseño de este prototipo.

La revisión de las herramientas existentes actualmente permiten evidenciar la emergente industria de las aplicaciones que buscan recoger y presentar los contenidos de las redes sociales con variados objetivos entre los que se encuentran: monitoreo de opiniones sobre una marca, lectura resumida de las publicaciones de los contactos de un usuario en las redes sociales, mostrar información de tendencia, búsqueda de tweets o comprobación de la veracidad de la información.

Tras analizar los estudios referentes al geoposicionamiento de los usuarios fue posible evidenciar las dificultades respecto a este tema: sólo cerca del 20 % de los usuarios completan el campo ubicación del perfil, el resto lo completan con lugares muy generales o ficticios. Esta dificultad acotó las expectativas del prototipo de poder privilegiar las fuentes geográficamente más cercanas al lugar del hecho noticioso. El método implementado en este trabajo logra relacionar sólo el 17,54 % de los usuarios totales con una provincia específica de Chile. Existen variadas técnicas utilizadas para mejorar este relacionamiento en los diversos estudios analizados [8] [51] pero debido a su complejidad se pretende abordar como desarrollo futuro.

Otra de las dificultades enfrentadas fueron los límites para la obtención de datos impuestos por la API de Twitter, estas limitaciones afectaron en el periodo de tiempo necesario para recolectar datos.

En la sección 2 pudimos profundizar en el interesante y activo debate sobre la influencia real de los procesos de *gatekeeping* en la subjetividad generada por el usuario al informarse de una noticia. Otra arista interesante que comprende esta temática, es el ejercicio propio de informarse y el papel que debiesen jugar no sólo los medios de prensa, sino también, de manera personal cada una de las personas-usuarias que las consumen.

Una visión interesante al respecto a este ejercicio es la que entrega ¡cita LemondeDiplomatique¡/cita¡. En ese contexto, el presente prototipo facilitaría la tarea de informarse de las personas, al reunir en un mismo espacio comentarios, visiones y opiniones de distintos usuarios además de una lista de enlaces externos donde profundizar o complementar puntos de vistas recogidos.

La raíz del problema respecto a la dificultad de informarse que señala Juan radica en un problema de recursos, al contar con tiempo limitado se accede a un número limitado de fuentes de información. Este trabajo permite reducir este esfuerzo de recursos, centralizando y generando reportes sobre tópicos y hechos noticiosos completos, categorizados y con procesos de

gatekeeping transparentes.

Este mismo aspecto (el del acceso a más de una misma fuente de información) abre un análisis sobre la arquitectura del prototipo desarrollado ¿no es acaso una debilidad importante que solo cuente con una fuente de información como es Twitter?

Aún cuando Twitter, es una de las pocas redes sociales que trabaja constantemente en políticas de transparencia (respecto a las solicitudes de información por parte de los gobiernos) y plantean abiertamente una postura de transparencia frente a estos asuntos ¹, existen precedentes de decisiones comerciales-estratégicas que poseen componentes de censura.

- **Intereses comerciales** como es el caso del bloqueo parcial a Meerkat² competencia directa y de alto grado de utilización, competitora de Periscope, empresa comprada recientemente por Twitter para realizar streaming de vídeo.
- **Políticas de uso**, como la denegación de acceso a la API para Politwoops[27], aplicación que hacía visible tweets borrados de políticos en más de 30 países. La cual Twitter justificó de la siguiente manera: Ímagínese: ¿Cómo sería de estresante- o incluso terrorífico-twittear si fuera irrevocable o inalterable? Ningún usuario es más merecedor de esa capacidad que otro. De hecho, la eliminación de un tweet es una expresión del usuario”.
- **Contexto y regulaciones culturales** como es el caso de los sistemas de filtros reactivos (sobre cuentas o tweets) que aplica Twitter para restricciones legales y culturales de los distintos países (escondiendo esos contenidos en sus respectivos países pero dejándolos disponibles en el resto del mundo) [74].

Considerando lo anterior, con una arquitectura que depende de una sola fuente de información el prototipo se expone a filtros de información ejecutados por esta fuente, que aún cuando sean transparentados, es complejo verificar su real impacto sino no existe posibilidad de acceder al código fuente en cuestión.

Otro aspecto interesante abordado en 2 es referente a la discusión sobre las razones de la existencia para los procesos de *gatekeeping*. Una parte importante de autores sostienen que

¹Creemos que el intercambio abierto de información puede tener un impacto global positivo. Para ello, es vital para nosotros (y otros servicios de Internet) para ser transparentes acerca de las solicitudes del gobierno para la información del usuario y de las solicitudes del gobierno para retener contenido de internet, el crecimiento de estas investigaciones pueden tener un efecto negativo grave en la libertad de expresión con implicaciones reales en la privacidad de las personas”[74]

²Aplicación que permite realizar streaming de vídeo directamente a los seguidores del usuario en Twitter

estos filtros son intencionados y diseñados con el objetivo de moldear la realidad que se transmite mientras que otro grupo sostiene que éstos no son intencionales sino necesarios y de origen netamente operativo. En este trabajo, aún cuando se evitaron aplicar deliberadamente filtros de contenidos (de naturaleza editorial o ideológica), fue necesario - debido a la gran cantidad de datos disponibles - generar clasificaciones y selecciones para extraer y presentar la información relevante, sin estos tratamientos (debido a la gran cantidad de tweets recogidos) éstos carecen de valor. Considerando esta situación, se considera que la única solución coherente a esta situación es transparentar los procesos de *gatekeeping* a sus usuarios, exponiendo de qué forma actúan y cómo se aplican, de esta manera los usuarios podrán verificar el real efecto que implican en un medio. A modo de metáfora, si tuviéramos la oportunidad de abrir las salas de prensas a miles de auditorías ciudadanas libres, éstas podrían verificar y corroborar las etapas de *gatekeeping* de dicha sala de prensa y validarlas para generar confianza.

El ejercicio de informarse sobre los hechos noticiosos es fundamental para la generación de opinión ciudadana, una herramienta como la presentada, contribuye a esta labor en cuanto facilita el acceso a informaciones difundidas por otros ciudadanos. Un aspecto interesante a analizar en este trabajo es el grado de pluralidad presente en las opiniones recogidas por el prototipo, si el escenario garantiza pluralidad equitativa este medio, sería una potente característica de herramienta informativa. Debido a su complejidad queda como trabajo futuro desarrollarlo.

REFERIRSE A LOS RESULTADOS ENCONTRADOS

7.2. Trabajo futuro

El presente trabajo presenta algunas componentes que podrían ser mejoradas para profundizar y mejorar aún más los resultados obtenidos. A continuación se presentan los distintos aspectos considerados para desarrollos futuros:

Referente al geoposicionamiento de los usuarios y la escasa tasa de llenado del campo 'ubicación' en sus perfiles se considera necesario como trabajo a futuro desarrollar enfoques con mejores resultados considerando no sólo datos relativos a los usuarios sino también de los distintos tweets (Como el abordado en [8] [36] [22] [35]). Un enfoque interesante es el de analizar el contenido de los tweets mediante la identificación de hashtags locales y palabras locales, fortaleciéndolo con temáticas exclusivas y delimitadas a dicha zona para identificar o

relacionar tweets con esa zona geográfica.

En cuanto al desempeño del prototipo se pueden realizar importantes mejoras. Una de las opciones más atractivas (relación rendimiento, escalabilidad y precio) es incorporar algunos de los Servicios Web de Amazon (AWS). AWS son un conjunto de servicios escalables tanto en costos como en capacidad, permitiendo el acceso a hardware de alto desempeño y la automatización de procesos a un costo accesible (POR QUE ES ACCESIBLE). Entre los servicios que ofrece se encuentran: cómputo global, almacenamiento, bases de datos, análisis, aplicaciones e implementación de servicios.

Los servicios que pueden contribuir directamente a mejorar el desempeño del prototipo son los siguientes:

- **Amazon RDS:** Proporciona un servicio seguro, escalable y simple de administrar para bases de datos en la nube. Proporcionando una capacidad rentable y de tamaño variable ante posibles crecimientos. AWS RDS proporciona seis motores de base de datos entre los que se encuentran MySQL.

Incorporar este servicio mejoraría sustancialmente el proceso actual de acceso a los datos para su análisis y eliminando las limitaciones de almacenamiento de datos. Su uso estaría destinado para almacenar los datos captados desde Twitter mediante la API y transferir conjuntos de datos a la base de datos cada cierta unidad de tiempo, para su almacenamiento permanente.

- **Amazon Kinesis:** Proporciona un servicio de procesamiento de datos en tiempo real a streaming de gran cantidad de datos. Amazon Kinesis puede capturar continuamente y almacenar terabytes de datos por hora a partir de cientos de fuentes de datos.

La incorporación de este servicio puede contribuir significativamente en la reducción de tiempos que toman las distintas fases del procesamiento de los datos para la creación de un nuevo tópico abriendo también una increíble oportunidad de re-estructuración completa del proceso de captación de tweets, migrando la captación desde la API REST de Twitter (con el análisis post del conjunto de tweets) a la captación de tweets desde la API STREAMING de Twitter (aplicando el análisis en tiempo real), aumentando considerablemente la frescura de la información ofrecida por el prototipo.

- **Amazon Elastic Compute Cloud (EC2):** Es un servicio web que proporciona capacidad de cálculo escalable en la nube. EC2 cuenta con una interfaz fácil de uso que entrega un control completo de los recursos informáticos, entregando la posibilidad de arrancar nuevas instancias de servidor en segundos, permitiendo escalar rápidamente la capacidad a medida que cambian las necesidades.

Estas posibilidades repercuten directamente en los tiempos de respuesta de las distintas fases del análisis del prototipo como la aplicación del clasificador, la recolección de enlaces, el orden de los distintos rankings entre otros, aumentando la capacidad de generación de los diversos tópicos.

En la sección previa referente a las conclusiones se plantea la importancia de la transparencia en los distintos procesamientos que se aplican a la información en la construcción de una noticia, es por ello, que un trabajo a futuro relevante es la habilitación de este prototipo para uso y la disposición ante la comunidad del código fuente desarrollado. La habilitación de este prototipo de un servidor de acceso público fue descartado como desarrollo de este trabajo debido a la inviabilidad económica de su mantención a mediano-largo plazo.

Otro aspecto relevante observado durante el desarrollo de este trabajo es el bajo volumen de documentación existente sobre hábitos de consumo de información y comportamiento para informarse en territorio nacional mediante internet o las redes sociales. Por lo cual, un interesante trabajo a futuro es perfilar y obtener información que permitan caracterizar a la población de usuarios de Twitter en Chile, esta información irá en directo beneficio para la comunidad de desarrolladores e investigadores sobre la materia en territorio nacional.

Bajo esta misma perspectiva, y considerando el principal enfoque de diseño de Storyful “vivir dentro de las comunidades de medios sociales, no para observar desde una distancia segura” sería interesante incorporar componentes participativas en el prototipo de tal manera que los mismos ciudadanos puedan contribuir a la divulgación y generación de información, en un esquema democrático y sin preferencias ni discriminaciones arbitrarias. En el último tiempo, los medios de prensa han incluido reportes de noticias ciudadanas de manera parcial (ya que sólo se limita a utilizar el material proporcionado por el reporte, sin incluir componentes sociales del usuario, comentarios de éste u información deducida en base a la interacción con el círculo humano con la intención de profundizar en la situación presentada), habilitando un

número de contacto donde se pueden enviar videos a través de Whatsapp o Twitter,

Capítulo 8

Anexo

Bibliografía

- [1] José Luis Martinez Albertos. La información en una sociedad industrial. *Tecnos*, page 119, 1981.
- [2] AmorTV. Sitio web de amor tv. <http://www.amortv.cl>. Accessed: 2015-11-07.
- [3] M. Becker. Die aktualität von onlineenzyklopädien – eine empirische analyse am beispiel wikipedia. 2012. Diploma Thesis University of Cologne.
- [4] Shayne Bowman and Chris Willis. *we media; media; citizens media; participatory media; blogging*. Colection ourmedia. The Media Center, American Press Institute, 2003.
- [5] W. Breed. *Social Control in the Newsroom: A Functional Analysis*. Bobbs-Merrill Reprint Series in the Social Sciences, S34. Bobbs-Merrill, 1955.
- [6] Axel Bruns. Gatewatching, not gatekeeping: Collaborative online news. *Media International Australia Incorporating Culture and Policy: quarterly journal of media research and resources*, 107:31–44, May 2003. This is the author-manuscript version of this paper. Please refer to the journal (link above) for access to the definitive, published version.
- [7] Directorio cartográfico de España y Latinoamérica. Directorio cartográfico de españa y latinoamérica. <http://www.dices.net/mapas/chile/mapa.php?nombre=Abarca&id=103>. Accessed: 2015-10-30.
- [8] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet: A content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 759–768, New York, NY, USA, 2010. ACM.

- [9] Biobio Chile. Buscador bio bio. <http://busca.biobiochile.cl/>. Accessed: 2015-10-24.
- [10] DeWitt Clinton, Mike Taylor, and many contributors. Github python-twitter. <https://github.com/bear/python-twitter>. Accessed: 2015-10-25.
- [11] Grupo Copesa. La tercera. <http://www.latercera.com/resultadoBusqueda.html?q=>. Accessed: 2015-10-24.
- [12] Oracle Corporation. Website of mysql workbench. <https://www.mysql.com/products/workbench/>. Accessed: 2015-10-25.
- [13] Anish Das Sarma, Atish Das Sarma, Sreenivas Gollapudi, and Rina Panigrahy. Ranking mechanisms in twitter-like forums. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 21–30, New York, NY, USA, 2010. ACM.
- [14] Subsecretaría de Desarrollo Regional y Administrativo. Códigos únicos territoriales actualizados. http://www.sinim.gov.cl/archivos/centro_descargas/modificacion_instructivo_pres_codigos.pdf. Accessed: 2015-10-25.
- [15] Asociación Nacional de la Prensa. Asociado a anp. <http://anp.cl/quienes-somos/asociados/>. Accessed: 2015-10-25.
- [16] Asociación Nacional de la Prensa. Website de asociación nacional de prensa. <http://www.anp.cl/>. Accessed: 2015-10-25.
- [17] Asociación de Radiodifusores de Chile. Radios socias de archi. <http://www.archi.cl/radios-de-chile.html>. Accessed: 2015-10-25.
- [18] Asociación de Radiodifusores de Chile. Website de archi. <http://www.archi.cl/>. Accessed: 2015-10-25.
- [19] Asociación Nacional de radios comunitarias y ciudadanas de Chile. Asociados a anarcich a.g. <http://radioscomunitariaschile.cl/Oficial/Copied-Red-Nacional-de-radios-Comunitarias.php> <http://>

- radioscomunitariaschile.cl/Oficial/Copied-RED-REGIONAL.php. Accessed: 2015-10-25.
- [20] Asociación Nacional de radios comunitarias y ciudadanas de Chile. Website de anarcich a.g. <http://radioscomunitariaschile.cl/>. Accessed: 2015-10-25.
- [21] Anlei Dong, Ruiqiang Zhang, Pranam Kolari, Jing Bai, Fernando Diaz, Yi Chang, Zhaohui Zheng, and Hongyuan Zha. Time is of the essence: Improving recency ranking using twitter data. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 331–340, New York, NY, USA, 2010. ACM.
- [22] Mark Dredze, Michael J. Paul, Shane Bergsma, and Hieu Tran. Carmen: A twitter geolocation system with applications to public health.
- [23] ElementTree. Elementtree overview. <http://effbot.org/zone/element-index.htm>. Accessed: 2015-10-25.
- [24] Facebook. The open graph protocol. <http://ogp.me/>. Accessed: 2015-10-31.
- [25] Flipboard. Website corporative of flipboard. <https://flipboard.com/>. Accessed: 2015-10-18.
- [26] Django Software Foundation. Website of django. <https://www.djangoproject.com/>. Accessed: 2015-10-25.
- [27] Open State Foundation. Twitter cuts off diplotwoops and politwoops in all remaining 30 countries, 2015. <http://www.openstate.eu/2015/08/twitter-cuts-off-diplotwoops-and-politwoops-in-all-remaining-30-countries/>.
- [28] Python Software Foundation. Documentación de python. <https://docs.python.org>. Accessed: 2015-10-25.
- [29] Python Software Foundation. Pypi - the python package index. <https://pypi.python.org/pypi>. Accessed: 2015-10-25.

- [30] Hauke Fuehres, Peter A. Gloor, Michael Henninger, Reto Kleeb, and Keiichi Nemoto. Galaxysearch - discovering the knowledge of many by using wikipedia as a meta-searchindex. *CoRR*, abs/1204.3375, 2012.
- [31] Tobias Futterer, Peter A. Gloor, Tushar Malhotra, Harrison Mfula, Karsten Packmohr, and Stefan Schultheiss. Wikipulse - a news-portal based on wikipedia. 2013.
- [32] H.J. Gans. *Deciding What's News: A Study of CBS Evening News, NBC Nightly News, Newsweek, and Time*. Medill School of Journalism Visions of the American Press Series. Northwestern University Press, 1979.
- [33] Geofeedia. Website corporative of geofeedia. <https://geofeedia.com/>. Accessed: 2015-10-18.
- [34] L. Gomis. *Teoria Del Periodismo: Como Se Forma el Presente*. Paidos Comunicacion. Ediciones Paidos Iberica, S.A., 1991.
- [35] Eduardo Graells-Garrido and Barbara Poblete. #santiago is not #chile, or is it? A model to normalize social media impact. *CoRR*, abs/1309.1785, 2013.
- [36] Mark Graham, Scott A. Hale, and Devin Gaffney. Where in the world are you? geolocation and language identification in twitter. *The Professional Geographer*, 66(4):568–578, 2014.
- [37] Antti Haapala. Python-levenshtein. <https://pypi.python.org/pypi/python-Levenshtein/>. Accessed: 2015-10-25.
- [38] T. Harcup. *Journalism: Principles and Practice*. SAGE Publications, 2004.
- [39] Alfred Hermida. Twittering the news. *Journalism Practice*, 4(3):297–308, 2010.
- [40] Alfred Hermida. Twittering the news: The emergence of ambient journalism. *Journalism Practice*, 4(3):297–308, 2010.
- [41] 24 horas. Buscador. <http://search.24horas.cl/search/?q=>. Accessed: 2015-10-24.

- [42] A.L. Hughes and L. Palen. Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management*, 6(3):248–260, 2009.
- [43] Ted. J. Smith III. La mordedura del perro guardian. *revista FACETAS*, 1991.
- [44] Twitter Inc. Documentation twitter developers. <https://dev.twitter.com/overview/documentation>. Accessed: 2015-10-25.
- [45] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, WebKDD/SNA-KDD '07, pages 56–65, New York, NY, USA, 2007. ACM.
- [46] Andrés Azócar Z. y Andrés Scherman T. Julián González U. Encuesta de caracterización de usuarios de twitter en chile. <http://www.prensaefcl.udp.cl/usuariotwitter.html> <http://www.prensaefcl.udp.cl/usuariostwitter.pdf>. Accessed: 2015-11-07.
- [47] Infographic Labs. Twitter 2012. <http://infographiclabs.com/news/twitter-2012/#more-722>. Accessed: 2015-11-07.
- [48] Harold D. Lasswell. The structure and function of communication in society. pages 215–228, 2007.
- [49] K. Lewin. *Field theory in social science: selected theoretical papers*. Social science paperbacks. Harper, 1951.
- [50] W. Lippmann. *Public Opinion*. Harcourt, Brace, 1922.
- [51] Jeffrey McGee, James A. Caverlee, and Zhiyuan Cheng. A geographic study of tie strength in social media. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 2333–2336, New York, NY, USA, 2011. ACM.
- [52] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.

- [53] Paper.li. Website corporative of paper.li. <http://paper.li/>. Accessed: 2015-10-18.
- [54] Marco Pennacchiotti and Ana-Maria Popescu. A machine learning approach to twitter user classification. In Lada A. Adamic, Ricardo A. Baeza-Yates, and Scott Counts, editors, *ICWSM*. The AAAI Press, 2011.
- [55] Owen Phelan, Kevin McCarthy, and Barry Smyth. Using twitter to recommend real-time topical news. In *Proceedings of the Third ACM Conference on Recommender Systems*, RecSys '09, pages 385–388, New York, NY, USA, 2009. ACM.
- [56] Jodok Batlogg edleaf glen.tregoning Brad Choate Jim Cortez Jason Lemoine Thomas Dyson Robert Laquey Hameedullah Khan Mike Taylor DeWitt Clinton Pierre-Jean Coudert, Omar Kilani and the rest of the python-twitter mailing list. Google project hosting for python-twitter. <https://code.google.com/p/python-twitter/>. Read Only Accessed: 2015-10-25.
- [57] Puroperiodismo. Menciones en twitter a medios chilenos de noticias. http://www.puroperiodismo.cl/?page_id=24464. Accessed: 2015-11-07.
- [58] Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. Classifying latent user attributes in twitter. In *Proceedings of the 2Nd International Workshop on Search and Mining User-generated Contents*, SMUC '10, pages 37–44, New York, NY, USA, 2010. ACM.
- [59] Stephen D. Reese and Jane Ballinger. The roots of a sociology of news: remembering mr. gates and social control in the newsroom. 2007.
- [60] J.F. Revel. *El conocimiento inútil*. Coleccion al filo del tiempo. Planeta, 1990.
- [61] Ramón Salaverría. *Redacción periodística en Internet*. Eunsa, 2005.
- [62] Semiocast SAS. Brazil becomes 2nd country on twitter, japan 3rd netherlands most active country. http://semiocast.com/publications/2012_01_31_Brazil_becomes_2nd_country_on_Twitter_superseds_Japan. Accessed: 2015-11-07.
- [63] Clay Shirky. *Here Comes Everybody*. Penguin Books, London, 2008.

- [64] Jetbrains s.r.o. Website of pycharm. <https://www.jetbrains.com/pycharm/>. Accessed: 2015-10-25.
- [65] Ian Bicking Holger Joukl Simon Sapin Marc-Antoine Parent Olivier Grisel Kasimier Buchcik Florian Wagner Emil Kroymann Paul Everitt Victor Ng Robert Kern Andreas Pakulat David Sankel Marcin Kasperski Sidnei da Silva Stefan Behnel, Martijn Faassen and Pascal Oberndörfer. Documentation lxml. <http://lxml.de/>. Accessed: 2015-10-25.
- [66] Matthew Honnibal Roman Yankovsky David Karesh Evan Dempsey Wesley Childs Jeff Schnurr Adel Qalieh Lage Ragnarsson Steven Loria, Pete Keen and Jonathon Coe. Documentation of textblob. <http://textblob.readthedocs.org/en/latest/index.html>. Accessed: 2015-10-25.
- [67] Storyful.blog. Making lists: How to fine-tune your twitter antennae. <http://blog.storyful.com/2012/11/05/making-lists-how-to-fine-tune-your-twitter-antennae/#>. ViQcn36rTIU. Accessed: 2015-10-18.
- [68] Storyful.blog. Storyful tips and tools: Locating videos on a map. <http://blog.storyful.com/2012/09/25/storyful-tips-and-tools-locating-videos-on-a-map/#>. ViQhJX6rTIU. Accessed: 2015-10-18.
- [69] Summify. Website corporative of summify. <http://summify.com/>. Accessed: 2015-10-18.
- [70] DarTar Tbayer, Hfordsa and Romanesco. Quantifying quality collaboration patterns, systemic bias, pov pushing, the impact of news events, and editors' reputation, 2011. https://en.wikipedia.org/wiki/Wikipedia:Wikipedia_Signpost/Single/2011-11-28#Recent_research.
- [71] The Tweeted Times. Website corporative of the tweeted times. <http://tweetedtimes.com/>. Accessed: 2015-10-18.

- [72] Twitter. Website corporative of tweetdeck. <https://tweetdeck.twitter.com/>. Accessed: 2015-10-18.
- [73] Inc. Twitter. Acerca de la empresa. <https://about.twitter.com/es/company>. Accessed: 2015-11-07.
- [74] Inc. Twitter. Tweets still must flow. <https://blog.twitter.com/2012/tweets-still-must-flow>. Accessed: 2015-11-07.
- [75] Inc. Twitter. Twitter for newsrooms and journalists. <https://media.twitter.com/best-practice/for-newsrooms-and-journalists>. Accessed: 2015-11-10.
- [76] Twitterfall. Website corporative of storyful. <http://storyful.com>. Accessed: 2015-10-18.
- [77] Ibrahim Uysal and W. Bruce Croft. User oriented tweet ranking: a filtering approach to microblogs. In Craig Macdonald, Iadh Ounis, and Ian Ruthven, editors, *CIKM*, pages 2261–2264. ACM, 2011.
- [78] K. Wahl-Jorgensen and T. Hanitzsch. *The Handbook of Journalism Studies*. ICA Handbook Series. Taylor & Francis, 2008.
- [79] D. M. White. *The gate keeper: A case study in the selection of news*. Journalism Quarterly. Bobbs-Merrill, 1950.
- [80] Arturo Arriagada y Patricio Navia. *Intermedios, medios de comunicación y democracia en Chile*. Ediciones Universidad Diego Portales. Ediciones Universidad Diego Portales, 2013.
- [81] Elena Real Rodríguez y Sergio Príncipe Hermoso y Pinar Agudiez Calvo. Periodismo ciudadano versus periodismo profesional: ¿somos todos periodistas? *Estudios sobre el Mensaje Periodístico*, 13, 2007.
- [82] Yuto Yamaguchi, Tsubasa Takahashi, Toshiyuki Amagasa, and Hiroyuki Kitagawa. TuRank: Twitter user ranking based on user-tweet graph analysis. In *Proceedings of the*

11th International Conference on Web Information Systems Engineering, WISE'10, pages 240–253, Berlin, Heidelberg, 2010. Springer-Verlag.