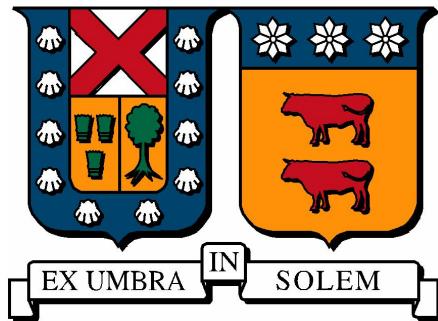


UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INFORMÁTICA
VALPARAÍSO, CHILE



**CREACIÓN DE PROTOTIPO PARA LA GENERACIÓN DE
REPORTES A PARTIR DE TWITTER**

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	VII
Índice Algoritmos	VIII
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	15
3.1.1.3. Mecanismo de Ranking en Twitter como foro	18
3.1.1.4. Twitter para la recomendación de noticias	19
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	22
3.1.2. Geolocalización de usuarios	24
3.2. Herramientas y plataformas relacionadas	27
3.2.1. Geofeedia	27
3.2.2. Paper.li	28
3.2.3. The Tweeted Times	29
3.2.4. FlipBoard	30
3.2.5. Summify	31
3.2.6. TwitterFall	33
3.2.7. Storyful	34
3.2.8. Wikipulse	36
3.2.8.1. Algoritmo de selección de noticias	37
3.2.8.2. Análisis de los datos	38
4. Definición de la solución	40
4.1. Objetivos	41
4.1.1. Objetivo principal	41
4.1.2. Objetivos Secundarios	41
5. Propuesta	42
5.1. Arquitectura de la solución	42
5.2. Plataformas y herramientas utilizadas	43
5.3. Implementación del prototipo	47
5.4. Características del servidor	47
5.5. Modelo de Datos	48
5.5.1. Análisis de la línea editorial del medio objetivo	48
5.5.2. Geolocalización de usuarios	50
5.5.2.1. Prueba valores distancia Levenshtein para identificar ubicación	50
5.5.3. Captación de usuarios	52
5.5.3.1. Captura de los medios de prensa (MP)	52
5.5.3.2. Captura followers de los medios de prensa (FMP)	53

5.5.4.	Captura de tweets	56
5.5.5.	Procesamiento de los tweets	57
5.5.5.1.	Definición del tópico	57
5.5.5.2.	Obtención del conjunto de tweets relacionados al tópico . . .	58
5.5.5.3.	Depuración de conjunto de tweets relacionados al tópico . . .	58
5.5.5.4.	Orden geográfico	59
5.5.5.5.	Orden de relevancia	60
5.5.5.6.	Panel de enlaces	61
5.5.5.7.	ON/OFF Medios de prensa	62
6. Evaluación y discusión		63
6.1.	Vistas del prototipo	63
6.2.	Caracterización de la población de datos capturados	67
6.3.	Resultados	72
6.3.1.	Referencias a noticias	73
6.3.1.1.	Medio de prensa Modelo	73
6.3.1.2.	Referencia a noticias	74
6.3.2.	Caso de prueba: El aborto	78
6.3.2.1.	Procesamiento del tópico	78
6.3.2.2.	Ánálisis de muestra representativa	79
6.3.2.3.	Ánálisis de contenido	79
6.3.2.4.	Ánálisis temporal	81
6.3.2.5.	Ánálisis de re-tweet	82
6.3.2.6.	Ánálisis geográfico	84
7. Conclusiones		86
7.1.	Conclusiones	86
7.1.1.	Conclusiones técnicas	86
7.1.2.	Consideraciones y discusión sobre las conclusiones	88
7.2.	Trabajo futuro	90

8. Anexo	93
8.1. Cotización en Amazon	93
8.2. Cuentas en Twitter de los medios de prensa	93
Bibliografía	107

Índice de figuras

2.1fig	Representación del proceso de gatekeeper [80]	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.	28
3.2fig	Imagen del panel de noticias de geofeedia, donde cada columna recoge los feeds para ubicaciones geográficas distintas.	28
3.3fig	Vista principal del periodico Paper.li de un usuario de la plataforma	29
3.4fig	Vista de fotografías del periodico Paper.li de un usuario de la plataforma	29
3.5fig	Vista principal de un periodico creado en The Tweeted Times	30
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	31
3.7fig	Vista de la recopilación de Summify	33
3.8fig	Vista de la recopilación de Twitterfall	34
3.9fig	Vista principal de Storyful	35
5.1fig	Diagrama conceptual de la arquitectura de la solución	43
5.2fig	Modelo de base de datos	48
5.3fig	Mapa con los usuarios por ubicación geográfica	51
5.4fig	Diagrama conceptual con las etapas para la captación de usuarios.	52
6.1fig	Vista nueva búsqueda tópico	63
6.2fig	Vista previa resultados de la búsqueda	64
6.3fig	Vista de tópicos	65

6.4fig	Vista de orden geográfico	65
6.5fig	Vista de orden por relevancia	66
6.6fig	Vista de orden temporal	66
6.7fig	Vista de links	67
6.8fig	Vista de tópicos	67
6.9fig	Distribución de tweets por hora de emisión	68
6.10fig	Distribución porcentual de tweets por hora de emisión según región	69
6.11fig	Tweet que generó nuevo record del mensaje más re-tweeteado.	70
6.12fig	Pneedimiento para determinar si un tweet se refiere a un hecho noticioso	75
6.13fig	Pneedimiento para verificar si el hecho noticioso haya generado una nota del medio modelo	76
6.14fig	Distribución de cantidad de tweets por día	81
6.15fig	Distribución de cantidad de tweets por periodos de quince días	81
6.16fig	Cantidad de re-tweets por quincena	82
6.17fig	Distribución geográfica de los usuarios	85

Índice de Algoritmos

1algo	Obtención de las palabras más frecuentes del timeline de un conjunto de tweets	49
2algo	Reconocimiento de ubicación del usuario mediante Levenshtein	50
3algo	Construcción lista de medios	53
4algo	Captura de usuarios	54
5algo	Gathering Followers	55
6algo	Gathering Followers con mejora	55
7algo	Algoritmo para la captura de tweets.	56
8algo	Obtención del conjunto de tweets relacionados al tópico	58
9algo	Clasificador Bayes-Naive para determinar si es miembro o no del tópico.	59
10algo	Origen Geográfico	60
11algo	Origen Relevancia	61
12algo	Obtención de enlaces externos contenidos en los tweets	61
13algo	Dominio para determinar si un tweet se refiere o no, a un hecho noticioso	76
14algo	Dominio para verificar si el hecho noticioso haya generado una nota del medio modelo.	77

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 [75] y presentando en el 2012 presentaba el impresionante ritmo de crecimiento de 11 nuevas cuentas de Twitter por segundo [49].

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 un estudio del uso de Twitter [47], Java 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 [47] 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 hace seguimiento de 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). 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.

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.

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 según un estudio realizado por Semiocast SAS [65].

Según el perfil de uso de Twitter en Chile desarrollado en [48] 76 % las y los chilenos utilizan Twitter varias veces al día, 16 % al menos utilizan Twitter 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 hrs. y 22:00 hrs. con un 73 % de las usuarias y usuarios.

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.[66]

El periodismo es un método de investigación y es 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. [40]

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.

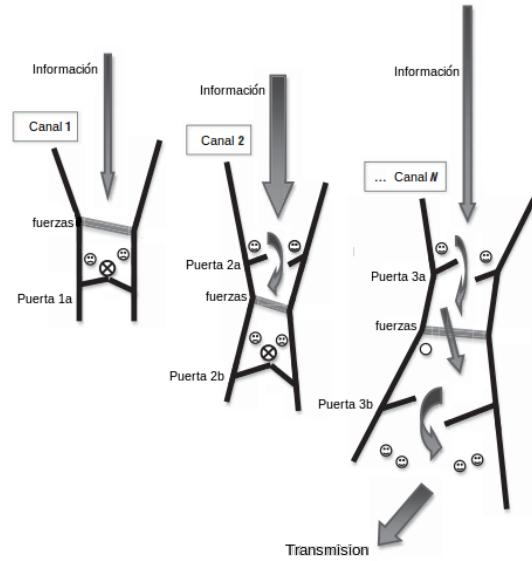


Figura 2.1: Representación del proceso de gatekeeper [80]

La figura 2.1 ilustra el proceso de *gatekeeping* donde se muestran tres canales, muchos elementos de información y 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. Los elementos de información 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. Las fuerzas negativas detienen el progreso de algunos elementos a través de los canales. El elemento final presentado como el resultado del proceso no sólo es el resultado de la selección, sino también 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[81] 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 [6] 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 [62] 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.

En [33] Gans 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 [51] 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 [52] 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 [50] 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 [45] explica precisamente una de estas tres funciones capitales, la de vigilancia del entorno 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

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

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 [63] critica la tendencia a priori que inducen los distintos medios de prensa en las distintas noticias². Real, Agudiez, Príncipe en [83] 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 [35] 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 los comunican y los efectos de ambas acciones son incomparables³. Para este autor el accionar

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 radiofónicas 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”

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”

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 [41] 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*” [5] 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. 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.

⁴“Nuestro estudio indica que la selección, o filtro, es el proceso de *gatekeeping* más cerrado para los usuarios, y creemos que seguirá siéndolo” [41]

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 [5] 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 [7] 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 [41].

Arriagada y Navia en [82] 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 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.”

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 [42] 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” [77].

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).[64]

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 usuarias y usuarios que forman hoy en día la red social de Twitter, la clasificación tanto de cuentas como de contenidos se ha convertido en una creciente 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

Pennacchiotti y Popescu en [56] desarrollan una clasificación del perfil de las cuentas considerando principalmente cuatro aspectos diferentes:

1. *Perfil del usuario*: Información básica de la cuenta como el nombre de usuario, la localización, una pequeña biografía además del número de seguidores y seguidoras, el número de personas a las que sigue y el número de tweets.
2. *Comportamiento para escribir tweets*: Conjunto de métricas de las interacciones existente entre la red social y el usuario o usuaria: el número promedio de tweets por minuto, número de respuestas, entre otras.

3. *Contenido lingüístico de los tweets:* Encapsula los temas principales de interés del usuario, así como su uso de léxico, este análisis mediante el uso de *palabras prototípicas*¹ permite la clasificación de los usuarios y usuarias según su estilo de escritura tales como textos oficiales, blogs, conversaciones o traducciones.
4. *Información de Twitter:* Estas características exploran las relaciones sociales establecidas por la o el usuario con los demás que él o ella siguen, a quien le responde o que personas re-tweetea. Pennacchiotti y Popescu indican que 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).

Respecto al análisis del *Perfil del usuario* Pennacchiotti y Popescu tras analizar un corpus de 14 millones de cuentas, identifican que sólo el 48 % provee una biografía corta y 80 % una ubicación, de cuya información se intentaron determinar el género del usuario o usuaria y su etnicidad pero los resultados obtenidos fueron de muy baja calidad.

Mediante una muestra 15000 cuentas de forma aleatoria y un conjunto de editores y editoras a quienes se pidió identificar la etnicidad y el género a partir de la imagen del avatar de Twitter. Se obtiene que menos del 50 % de las imágenes se correlaciona de manera clara con alguna etnia mientras que el 57 % se correspondía con algún género. Se identifica también que las imágenes podían ser engañosas: en el 20 % de los casos la imagen no corresponde al dueño o dueña de la cuenta sino de una celebridad o de otra persona.

Con esta información estadística el estudio concluye que los campos del perfil del usuario no contienen suficiente información ni de buena calidad para ser utilizada para una clasificación.

Referente al *Contenido lingüístico de los tweets* se realiza un análisis de los hashtag prototípicos, basados en la hipótesis de que si los usuarios o usuarias 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.

¹ El análisis se realiza mediante el uso de las cuales son expresiones típicas de las personas pertenecientes a una clase específica, así como frases relacionadas a intereses típicos de dicha clase

También se realiza un análisis de sentimientos a los distintos 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. Para este propósito, se recoge 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.

Las clases específicas estudiadas se refieren a: afiliación política, clasificación de si es o no seguidor de Starbucks y si los usuarios y usuarias pertenecen o no a la etnia afroamericana.

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 "tweeter".

'Para determinar si un usuario o usuaria pertenece a la clase *seguidor o seguidora potencial de Starbucks* se identifica que la *información del perfil y análisis lingüístico* son las características más útiles para este objetivo.

Se concluye además que la relación entre las y los seguidores y amigas y amigos es también una característica relevante por sí sola para determinar la clase de seguidores o seguidoras, 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 lingüísticas han sido ampliamente adoptadas por otros grupos haciendo visible que 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 a tarea de clasificación puede ser ayudada por información del perfil.

3.1.1.2. Ranking de enlaces compartidos en Twitter

En [23] Dong utiliza como fuente de información fresca los contenidos y enlaces compartidos en Twitter para búsquedas web en tiempo real. Dong identifica que en vista de la investigación de Hughes en [44] 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 enlaces que no son noticias y mejorar los resultados de la búsqueda).
- Los enlaces difundidos son publicados en base a las distintas prioridades personales de las y los usuarios, lo que aporta un interesante grado de diversidad.
- La red de Twitter permite realizar mediciones de autoridad a las y los creadores de tweets.
- Los tweets cuentan con metadatos relacionados que permiten clasificarlos e inferir en base a su relevancia.

Se consideran filtros en el procesamiento de enlaces referente al spam *Corresponden a enlaces propagados en Twitter referente a un producto o marca sin contenido relevante* basados en heurísticas a fines (como filtrar enlaces twitteadas por el mismo usuario más de dos veces o solo twitteadas por un mismo usuario).

Referentes a los enlaces y los tweets que los contenían se consideran las siguientes características:

- **Características textuales:** Se considera que las palabras que acompañan a una URL en un tweet pueden entregar información relevante sobre ésta. 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 [56]).
- **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 las y los usuarios, número promedio de tweets de las y los usuarios, numero promedio de las y los usuarios que retweetean cantidad de tweets que contienen la URL, promedio de usuarias y usuarios que responden los tweets que contienen la URL, promedio del número de usuarias y usuarios a las que siguen, 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 se basa en una máquina de aprendizaje de ranking (MLR) que considera 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 tweets y su relación con las palabras relacionadas, número de seguidoras y segidores del usuario con mayor grado de autoridad, número de usuarias y usuarios que re-tweetean la URL al usuario con mayor grado de autoridad, número promedio de usuarios y usuarias que re-tweetean los tweets que contienen la URL y número promedio de usuarios y usuarias 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 y se contrastaron con la opinión de cinco expertos en cinco grados de clasificación: perfecto, excelente, bueno, justo y malo. Se agrega además 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 cuentan son de la categoría: *sin sensibilidad de tiempo, reciente o de alguna manera reciente* al momento de su clasificación preservan su clasificación, en cambio,

cualquier momento 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 evaluaran una query como “con contenidos relevantes de última hora” 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 y que una 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.

Se identifica también que el porcentaje de edades de los enlaces clasificados 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 [15] 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 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 y usuarias 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, su evaluación depende principalmente del orden en que son presentados, 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, el cual funciona de la siguiente manera: cuando el usuario o usuaria envía un nuevo elemento (tweet, comentario o 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).

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 [57] utiliza Twitter y fuentes RSS para este cometido.

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 de la usuaria o usuario.
- *Index Lucene²:* Responsable de la indexación y la minería de la información obtenida.

²Lucene es una biblioteca de búsqueda de texto completo extremadamente rica y poderosa escrita en Java

- *Recomendación:* Encargado de generar una lista clasificada de historias RSS basado en la co-ocurrencia de términos populares dentro de los post y gustos expresados 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. De esta forma, se crea un ranking de temas y artículos a sugerir (ordenados entre estos mismos según su grado de relevancia).

Por otra parte el usuario puede elegir tres estrategias distintas de recomendación:

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

Se evalúa la aceptación de las recomendaciones realizadas por este sistema mediante el criterio de un pequeño grupo de 10 participantes, en un período de 5 días y se cuantifica la cantidad de clicks recibidos por noticia recomendada por el sistema. Cada participante configuró el sistema con su información correspondiente.

Al analizar los resultados se obtiene una clara diferencia en el comportamiento de las y los usuarios cuando se comparan las estrategias basadas en Twitter a la basada en el contenido predeterminado. 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 es el *ranking de las y los amigos* en comparación a las recomendaciones del *ranking público*, resultado que se contradice con un cuestionario posterior realizado, donde un 67 % de las y los usuarios indican su preferencia por *ranking público* mientras que sólo el 22 % señala su preferencia

por el *ranking de las y los amigos*. Ninguno de los participantes se muestra partidario de la 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

Ibrahim y Bruce en [79] plantean que la acción de re-tweetear para clasificar a los usuarios posee un gran valor: re-tweetear 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 la usuaria o usuario considera el tweet como información relevante. Basado en esta apreciación, se busca dar respuesta a la interrogante ¿Es posible clasificar a los usuarios basado en si retuitean 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 determinar la precisión con la que se puede clasificar los tweets como retweetable o no para un usuario específico. Se entrena el clasificador se utilizan cuatro grupos de rasgos:

- *Basado en el autor*: Se refiere a características deducidas a partir del perfil del usuario, relacionadas a qué tan activo es el usuario o usuaria y la autoridad que posee.: ¿Es el usuario un autor de élite³ ?, cantidad de followers, cantidad de personas a las que sigue, cantidad de tweets, edad del perfil⁴, tasa de tweets, cantidad de favoritos, ¿tiene descripción?, ¿su idioma es el inglés?.
- *Basado en los tweets*: Se refiere a características sintácticas del tweet, algunas de éstas dan implicaciones sobre que tan bien está escrito el tweet: la categoría, la audiencia y la popularidad del tweet.: La puntuación TF-IDF⁵), ¿contiene hashtag? ¿contiene urls?; Menciona

³élite 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 y 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.

⁴cantidad de días desde la creación de la cuenta

⁵TF-IDF es una estadística numérica que se pretende reflejar la importancia de una palabra es un documento, colección o corpus

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 y el largo promedio de los tweets.

- *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) y 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: Del tweet re-tuiteado ¿sigue el usuario al autor? ¿utiliza el hashtag del tópico relacionado en otros tweets? ¿se comparten enlaces con el tweet? ¿se menciona al usuario en el tweet en cuestión?

Para el conjunto de entrenamiento se realizan dos clases: los tweets re-tuiteados y los que no son re-tuiteados. Los resultados obtenidos señalan que ningún grupo de características arrojó resultados satisfactorios por si solos, entre las cuales la mejor característica 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 re-tuiteado 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?”, ”¿se menciona al usuario en el tweet en cuestión?”.

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

En [84] 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 las y los usuarios siguen de vuelta al usuario que comenzó a seguirlos por un acto de cortesía formal, por lo cual se considera 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-tweet).

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 dejan de seguirlo incluso si resulta que no transmite más información útil. Esto ocurre porque los usuarios no recuerdan a todas las usuarias y 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.

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 tiene un gran número de re-tweet y se propagan ampliamente. Por lo anterior, no es suficiente sólo contabilizar la cantidad de re-tweets para medir la autoridad de una usuaria o usuario, ya que los re-tweets conversacionales son menos relevantes para la puntuación de autoridad en cuestión.

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 cada nodo representa 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*⁶.

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

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

⁶ObjectRank es una extensión de PageRank[54] 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

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 re-tweet. 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.

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 re-tweets realizados por el mismo usuario o usuaria. Por otra parte, el ranking planteado es un eficaz sistema de puntuación para evaluar la autoridad de los usuarios de Twitter ya que evalúa a las usuarias y usuarios que no son seguidos por muchos usuarias y usuarios, pero sus tweets son re-tweeteados muchas veces, con una mayor posición mientras que a las usuarias y 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 aquellas usuarias y usuarios en los cuales la mayoría de sus tweets son conversaciones, son evaluados como completamente inútiles por el algoritmo.

3.1.2. Geolocalización de usuarios

Cheng en [9] identifica que la función de geolocalización en Twitter no es una función muy utilizada por las y los usuarios tras el análisis de una muestra aleatoria de más de un millón de usuarios, en la cual sólo el 21 % señala el campo *ubicación* del usuario como un nombre granular de una ciudad (por ejemplo: Los Angeles, CA) y sólo 5 % señala 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 [53] 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 las usuarias y usuarios con vínculos débiles.

Cheng propone un marco de trabajo para la localización de usuario, que cumpla con las siguientes características:

- 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 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 es una tarea difícil puesto que los tweets son inherentemente ruidosos, a menudo con expresiones coloquiales y vocabulario no estándar. No es obvio en lo absoluto que existan señales claras de ubicación incrustados en los tweets de algún usuario. Una usuaria o usuario puede tener intereses que abarcan múltiples ubicaciones y 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 en base al contenido de los tweets de una usuaria o usuario, caracterizándolo 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 kms. de sus ubicaciones reales (con una distancia de error promedio de 2.853 kms.). Se concluye que este enfoque no aporta 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 un población pequeña, tiene un conjunto escaso de palabras en sus tweets, lo que significa que la distribución de palabras por ciudades para estas ciudades, están 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.

Basado en la tercera metodología planteada por [9] Graells y Poblete en [36] desarrollan una técnica de posicionamiento geográfico para usuarios en territorio Chileno con intenciones de corroborar si la población virtual en Twitter es representativa respecto a su ubicación a la población física. Se utiliza un modelo de espacio vectorial, que dado el contenido de un tweet, permite (a través de un clasificador construido sobre modelos de lenguaje) establecer su ubicación geográfica.

Mediante un clasificador TD-IF 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. Finalmente se concluye 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 [34] es un servicio pionero en el control de origen de los medios de comunicación social mediante geolocalización, permite monitorear y realizar análisis a datos provenientes de las redes sociales, complementando la tradicional búsqueda de *keywords* que típicamente pierden información sobre la actividad social relacionada, ayudando a reducir el desorden de los medios sociales en tiempo real. 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 .ºlvidarusuarios (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 y exportar a través de la API en formatos CSV ATOM, GeoRSS, JSON, KML.

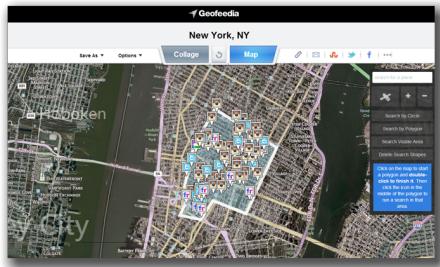


Figura 3.1: Imagen del mapa interactivo de geofeedia donde en una zona delimitada por el usuario se reciben todos los feeds de los medios sociales.

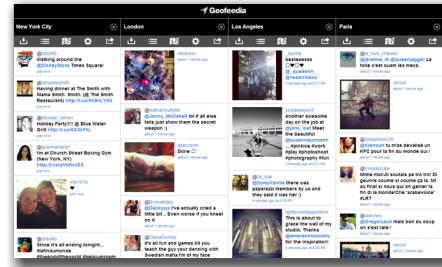


Figura 3.2: Imagen del panel de noticias de geofeedia, donde cada columna recoge los feeds para ubicaciones geográficas distintas.

3.2.2. Paper.li

Paper.li [55] 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, permitiendo fáciles combinaciones de las distintas fuentes de contenido, la priorización temática, la aplicación de filtros de contenidos y la configuración las fuentes de información.

El funcionamiento de Paper.li se basa en el hecho de que muchas usuarias y 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.

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 las noticias en las redes sociales: la reproducción de contenidos más frecuentes, los tópicos más consultados o lo más recomendados.

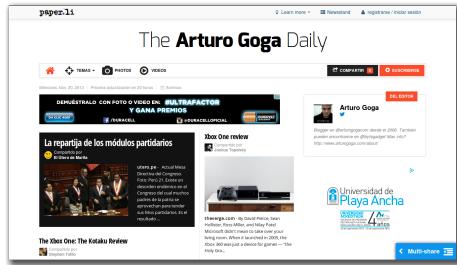


Figura 3.3: Vista principal del periódico Paper.li de un usuario de la plataforma

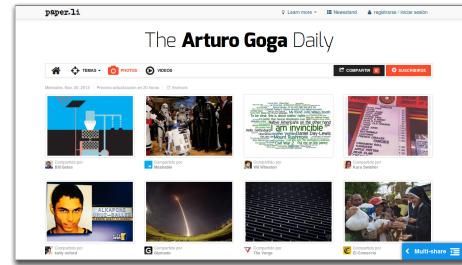


Figura 3.4: Vista de fotografías del periódico Paper.li de un usuario de la plataforma

3.2.3. The Tweeted Times

The Tweeted Times [74] 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, además de permitir explorar los periódicos de otros contactos en Twitter y de usuarios destacados de la plataforma. Para la organización de las publicaciones de acuerdo a su relevancia el algoritmo utilizado se basa principalmente en la repercusión de dichas publicaciones en Twitter (cantidad de re-tweet y número de favoritos).

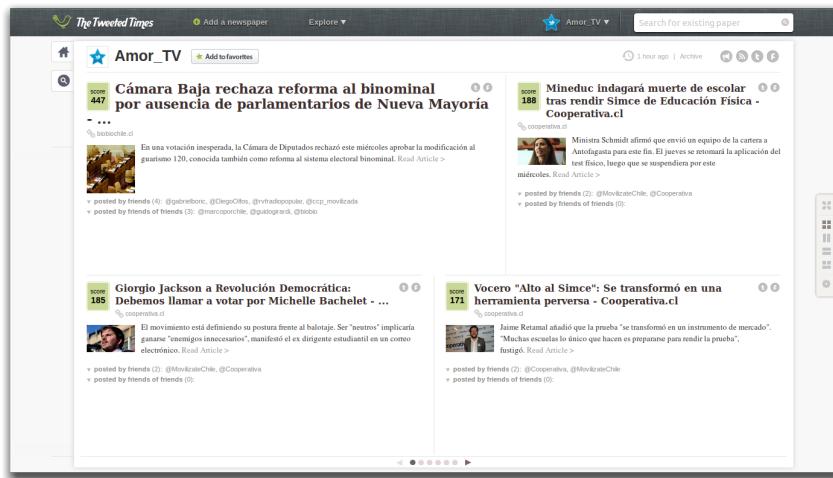


Figura 3.5: Vista principal de un periodico creado en The Tweeted Times

3.2.4. FlipBoard

Flipboard [27] 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, 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.

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 y usuarias contenidos seleccionados donde se incluyen las revistas y blog recomendados, fotografías y secciones especiales a noticias del día y temas de interés.

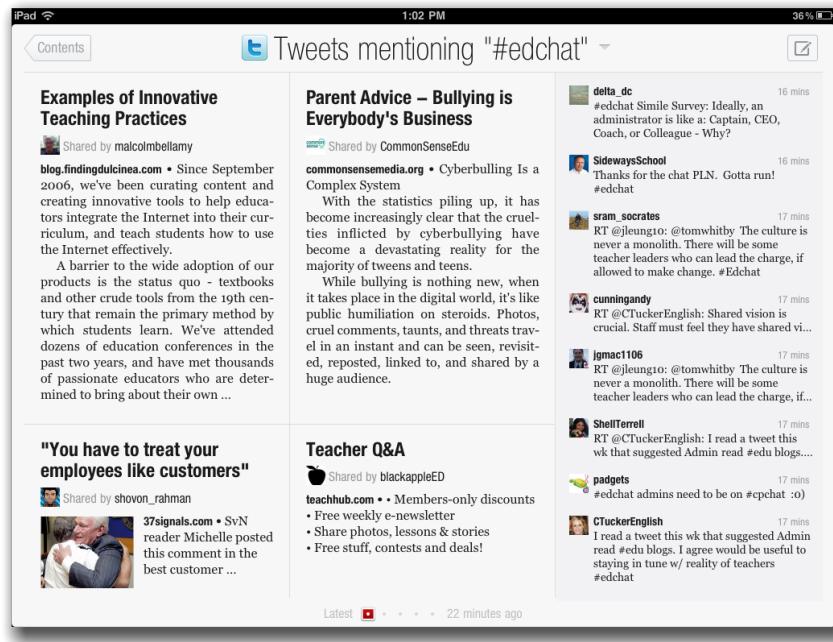


Figura 3.6: Vista del formato de revista en Flipboard para visualizar los feeds de Twitter, donde se resaltan los tweet que han tenido mayor repercusión en la red y los enlaces compartidos

3.2.5. Summify

Summify[71] 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, Facebook, 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 se refiere a lo costoso en tiempo que es mantenerse informado en las redes sociales⁷. Summify se presentaba como una alternativa productiva para mantener satisfactoriamente los vínculos sociales en la red.

Summify permitía la configuración de la cantidad de recopilaciones que se deseaba recibir el usuario (de una a cuatro diarias) y la hora de recopilación de la primera de éstas (las restantes

⁷...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 [72].

se realizaban en consecución en horas), era posible especificar la cantidad de artículos que poseían dichas recopilaciones (de dos a quince artículos), además de su privacidad (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 según su impacto (cantidad de veces compartido, cantidad de “Me gusta” o cantidad de re-tweeteados), prestando atención a las personas con las que el usuario interactua más.

El algoritmo consideraba también el impacto a nivel global pero con menor consideración que las recomendaciones de los contactos [4]. Otra cualidad del algoritmo es la capacidad de aprender en base al uso dado por la o el usuario, analizando las tendencias y preferencias de uso, observando si se pinchaban más historias de un usuario en particular, en un dominio concreto, si contienen palabras claves dentro del título o según 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 las y los usuarios que las habían compartido y al posicionar el puntero por encima de los avatares, el mensaje en el cual compartieron el artículo.

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).

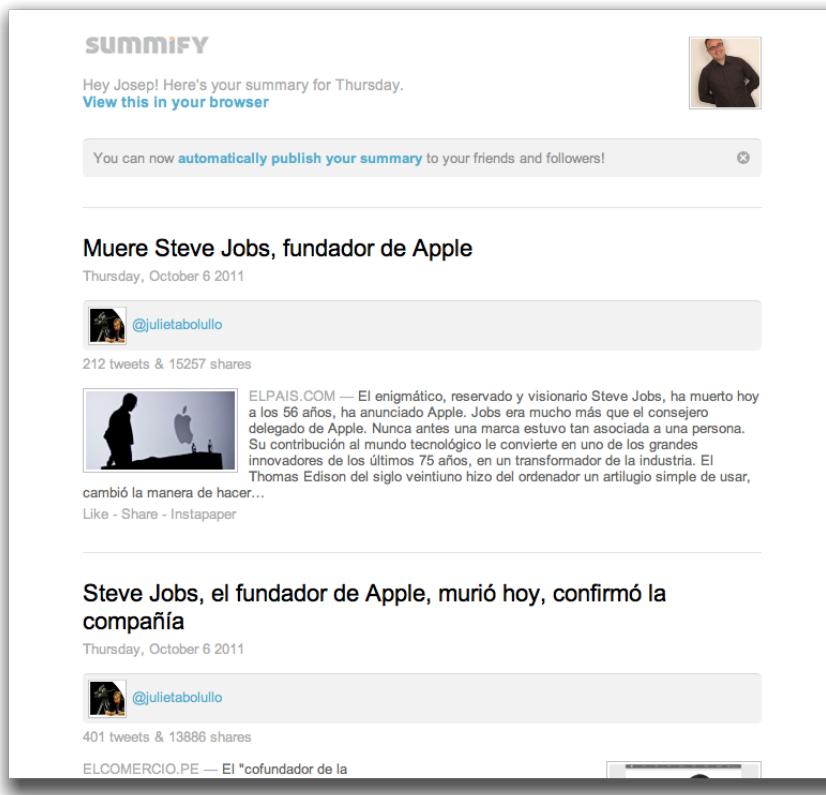


Figura 3.7: Vista de la recopilación de Summify

3.2.6. TwitterFall

Twitterfall [?] es un cliente web 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 tweet-osfera”. El nombre de la plataforma se relaciona con la forma de su interfaz, ya que se presenta como una cascada de tweets. éstos caen (hacia abajo de la pantalla) a medida que existen nuevas actualizaciones en tiempo real. 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 trending topic.

- Realizar búsquedas geolocalizadas.
- Realizar filtrados por idioma y excluir re-tweet.
- Almacenar las búsquedas favoritas.

TwitterFall permite también iniciar sesión en Twitter desde su interfaz y utilizarlo como cliente regular de Twitter para responder mensajes, realizar re-tweet, seguir a nuevas personas, etc. otorga además interesantes características como pre-visualizaciones en ventanas flotantes sobre enlaces contenidos en tweets (sin necesidad de ser redirido a ellos)



Figura 3.8: Vista de la recopilación de Twitterfall

3.2.7. Storyful

Fundada en Dublín con sólo tres empleados, Storyful [78] tiene su razón de ser en idear una manera de administrar las enormes cantidades de contenido que se comparten en las redes sociales, aplicando procesos y tecnología eficaz para ayudar a filtrar “noticias de ruido”. Storyful pretende dotar de herramientas a todos los medios de prensa que las demanden⁸.

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.

Storyful plantea que las y los periodistas deben pensar en su papel en un mundo cambiante, comenzando a utilizar nuevas herramientas inteligentes y rápidas, complementándolas con los

⁸ Tal como declaran en 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 de noticias en el mundo, incluyendo ABC News, Reuters y el New York Times, y plataformas sociales como YouTube.”

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. Aseguran 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, pero no sólo escuchar en ellos, sino participando directamente, de forma abierta y honesta con las voces más cercanamente relacionadas con el hecho⁹.

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.

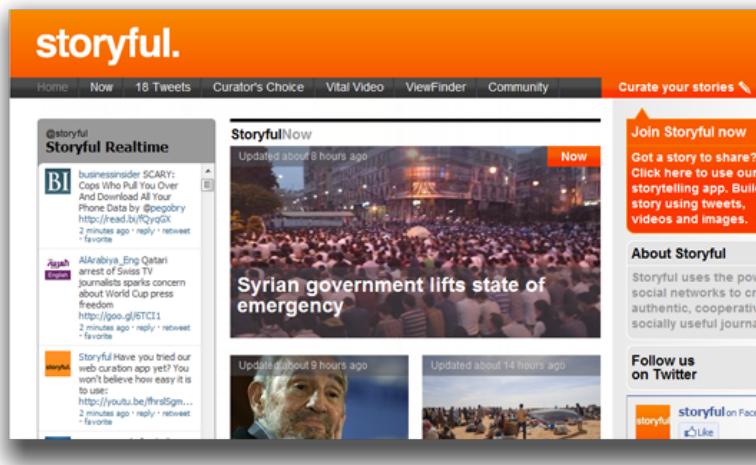


Figura 3.9: Vista principal de Storyful

⁹Tal como indican en su blog: ”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”<http://storyful.com/about/>

3.2.8. Wikipulse

Wikipulse es un medio de prensa [32] basado en el hecho de que la propia comunidad de editoras y 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 [73] 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 posee dos grandes componentes: Wikipedia actúa como fuente principal de artículos de noticias mediante artículos editados o publicados por la comunidad de Wikipedia, y la segunda componente está compuesta por la generación de noticias de Wikipulse de forma automática obtenidos desde Wikipedia agrupados en las distintas categorías (eventos corrientes, deporte, etc.) Estas componentes participan a través de los tres principales procesos de la plataforma:

1. Búsqueda de artículos relevantes en Wikipedia.
2. Reforma de los artículos en formato de noticias.
3. Presentación de los resultados.

Para el criterio de selección además del criterio obvio de los artículos editados más recientemente en Wikipedia, se emplea un algoritmo planteado en [31] donde 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 tipo de artículo de prensa. La comprobación de veracidad de los hechos del artículo 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 utilizando SMMRY¹⁰. 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.

3.2.8.1. Algoritmo de selección de noticias

El algoritmo se basa en tres grandes etapas:

1. **Creación del Conjunto de trabajo:** El algoritmo se ejecuta periódicamente y en cada ejecución procesa un conjunto de noticias, las cuales son llamadas *conjunto de trabajo*. El *conjunto de trabajo* contiene el conjunto de páginas recientemente editadas en Wikipedia. En primer lugar la página de metadatos, el título, la totalidad de las y 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 se realiza mediante la generación de un ranking por cada página, éstos son comparados con un valor de umbral t . Cada evaluación esta basada en consultas a la base de datos, ponderadas por un peso w que especifica la importancia de esa evaluación en el resultado final. Una página genera una noticia cuando se cumple que:

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

La mayor parte de las evaluaciones son ratios computados en orden por cada página para

¹⁰ SMMRY proporciona una manera eficiente de la comprensión de texto, que se realiza principalmente mediante la reducción del texto a las frases más importantes [67]

tener un conjunto predecible de números como resultado.

- Autor-con-Noticia Aumenta la importancia de las páginas de autores generadores-de-noticias (autores que ya han generado noticias). Se calcula como la cantidad de autores que editan la página dividido por todos los autores generadores de noticias.
- Autores-comunes Calcula la popularidad de una página mediante sus autores. Se calcula como la cantidad de autores y autoras editando la página dividido por todas y todos los autores de la base de datos.
- Dominio-experto Identifica las páginas importantes mediante el conteo del conjunto de expertos¹¹ que editan la página.
- Cambios-recientes Mide la actividad de edición en la página en comparación con todas las otras páginas del conjunto de trabajo.
- Relevancia La relevancia utiliza la evaluación de un webservice externo para determinar la popularidad de una página de Wikipedia.

3. Creación de noticias: La creación de noticias a partir de las páginas 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 *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 considerar una edición, esta es elegida en base a tres criterios:

- La edición debe ser hecha por un editor top¹².
- La edición debe ser hecha por un experto del dominio.
- La edición es más larga que 50 caracteres.

3.2.8.2. Análisis de los datos

Busca medir el rendimiento del seleccionador de noticias, de esta forma se decide comparar con algunas fuentes de información tradicionales como (CNN, Reuters, etc.). La idea principal

¹¹ Un experto del dominio es aquel autor en Wikipedia que edita páginas en la misma categoría que la página que se está evaluando.

¹² Una editora o editor top es aquella o aquel autor que está dentro de los “editores top”, lista que incluye las y los autores y expertas y expertos de dominio que tienen más ediciones para una categoría dada.

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 mas lento en términos de reporte.

Para comparar los resultados de Wikipulse se comparan los reportes con un medio convencional mediante el acceso a su RSS. Para comparar si se trataban del mismo contenido, se realiza un proceso de *match* entre ambos recursos de la siguiente manera: Las historias individuales obtenidas de ambos *feeds* se procesan para obtener una lista de palabras. Los distintos artículos (de ambos *feed*) se comparan mediante el *match* de las palabras claves. El cuantificador se calcula mediante una fracción entre el número total de *match* encontrados partido por el número total de *keywords* en cada uno de los artículos. Empíricamente se considera que un cuantificador superior a 0.33 generalmente indica un *match* fuerte.

Capítulo 4

Definición de la solución

Existe un creciente desconfianza por parte de la ciudadanía respecto a la repercusiones reales del 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 las y los ciudadanos para informarse de los eventos que ocurren en su entorno. Tambini en ¹ señala “El papel de *gatekeeper* de los medios de prensa tradicionales se ha debilitado y las personas construyen su propia estructura editorial eligiendo que 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 la televisión, 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 [39].

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 determinando evento, privilegiando aquellos reportes con menor cantidad de intermediadores y posibles *gatekeeper*, siendo estos, las y los observadores y las y los 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 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.

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.

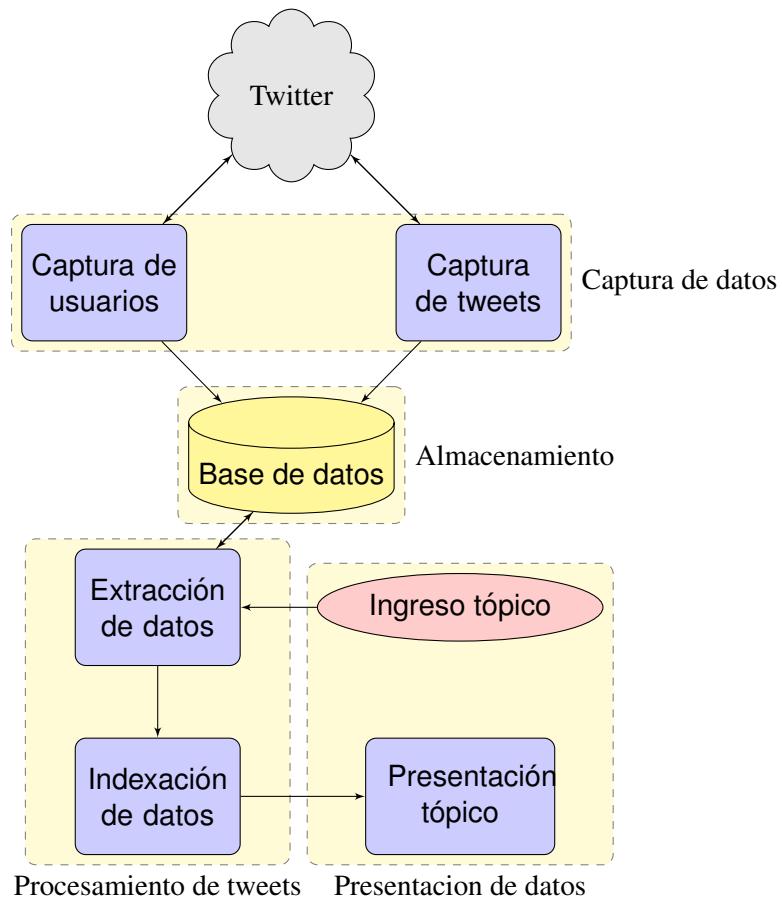


Figura 5.1: Diagrama conceptual de la arquitectura de la solución

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 con una sintaxis muy clara. Incorpora módulos, excepciones, interpretación dinámica y tipos de datos dinámicos de muy alto nivel, posee además variadas interfaces para muchas llamadas de sistema y bibliotecas así como de sistemas de ventanas y es extensible en C o C++.

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 extensiones desarrolladas por terceros [30], algunas de las cuales fueron ocupadas en este trabajo.

■ Librerías de Python utilizadas

• LMXL

LMXL [69] es un conjunto de herramientas que vinculan 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 [25]. La librería se encuentra bajo licencia BSD. Mientras que las librerías que extiende libxml2 y libxslt2 permiten su uso bajo licencia MIT.

• TextBlob

TextBlob [textblobWebsite](#) es una librería de Python para el procesamiento textual, posee una API simple para profundizar en las tareas del procesamiento del lenguaje natural (NLP en inglés) como etiquetar partes de un discurso, análisis emocional y traducciones. La librería posee una basta documentación y su licencia de uso permite el acceso, edición y uso de manera gratuita.

• Levenshtein

La extensión de Levenshtein [38] 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.
- Calcular la similitud de conjunto de strings.

Esta extensión posee licencia GNU.

■ MySQL

MySQL [13] 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 actualmente es usado por muchos sitios web populares como Wikipedia, Google, Facebook, Twitter, Flickr y Youtube.

■ Pycharm

Pycharm [68] 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 [28] 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 *plantillas*) 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**

MysqWorkbench [14] 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, auto-completado, 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 [46] 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 fácil uso se utiliza la librería Python Twitter [58] [11] que provee una interfaz en Python para la API de Twitter.

5.3. Implementación del prototipo

5.4. 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

5.5. Modelo de Datos

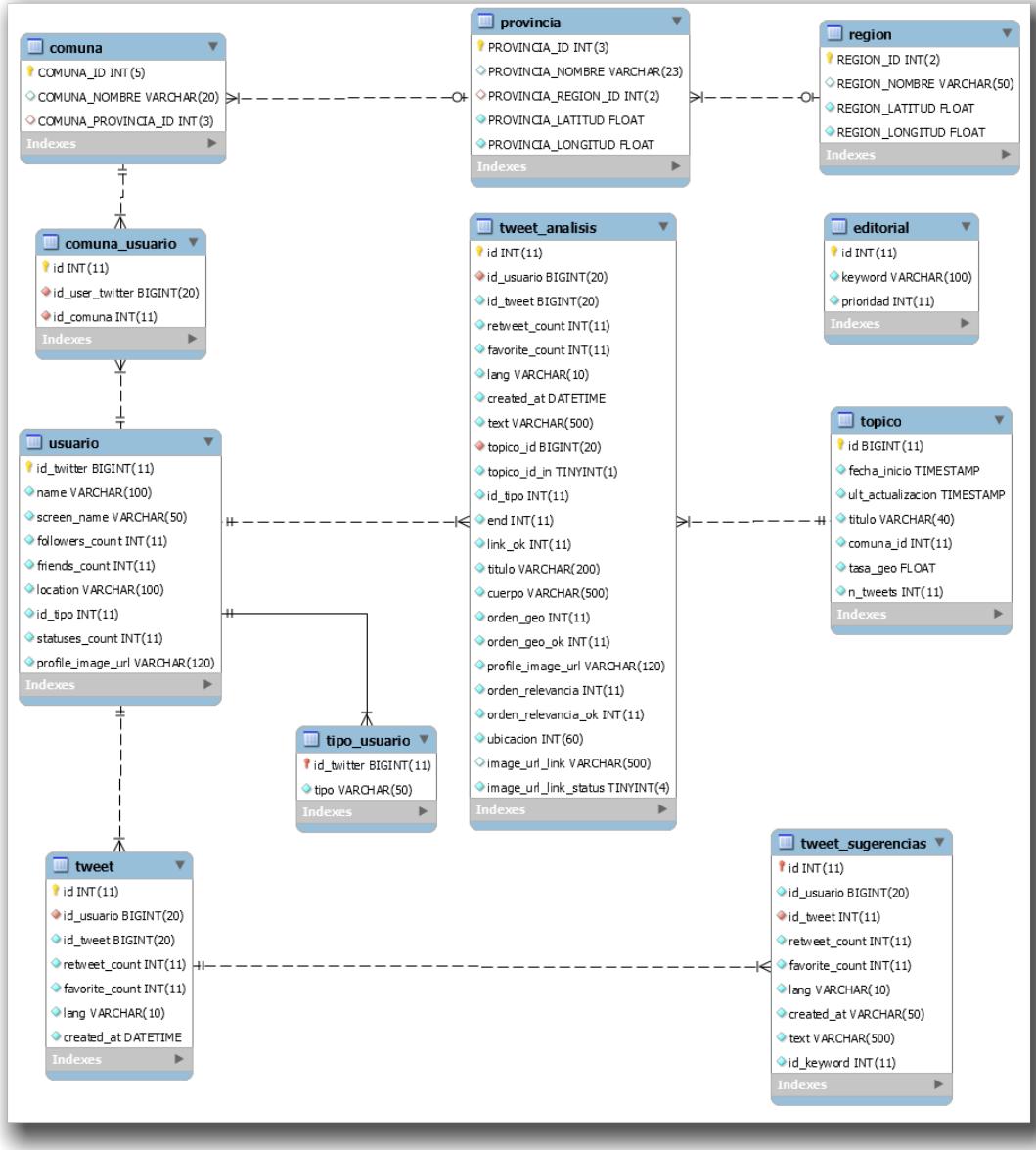


Figura 5.2: Modelo de base de datos

5.5.1. 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 de dicho medio. El proceso se realiza mediante el conteo de la frecuencia de las palabras contenidas en los corpus de los distintos

tweets, sin considerar las palabras vacías o stopword³ 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	Utsm
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

Se obtiene que las palabras con mayor frecuencia tienen relación con temáticas estudiantiles, manifestaciones, movimientos sociales y los lugares donde estos ocurren, que va acorde a la descripción del medio [2].

³Stopwords o palabras vacías es el nombre que reciben las palabras sin significado como artículos, pronombres, preposiciones, etc. que son filtradas antes o después del procesamiento de datos en lenguaje natural (texto)

5.5.2. Geolocalización de usuarios

El método implementado utiliza el concepto de la distancia de Levenshtein⁴ para determinar si el texto proporcionado por el usuario como ubicación corresponde o no a una comuna existente en Chile. La información relativa a las actuales comunas, 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 [16]. La posición GPS de cada una de las provincias se obtuvo mediante la ubicación proporcionada en [8].

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

```

5.5.2.1. Prueba valores distancia Levenshtein para identificar ubicación

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

Los resultados obtenidos fueron los siguientes:

D. Levenshtein	Nº match correctos	Nº match incorrectos	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 valores de distancia 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

⁴La distancia de Levenshtein corresponde a la cantidad de cambios necesarios en un string para transformarlo en un string objetivo.

que ocurren.

Se consideró que un error menor al 5 % es adecuado, considerando la cantidad de coincidencias correctas que aporta al conjunto. Por lo anterior se concluye que la distancia de Levenshtein a utilizar corresponde a 3. Considerando este parámetro 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). Éste porcentaje comparativamente es mayor que el 12 % encontrado por Cheng en [9] mediante su primera metodología explicada en 3.1.2.

Al disponer los distintos usuarios en un mapa utilizando la API de mapas de Google, se obtienen la siguiente visualización:

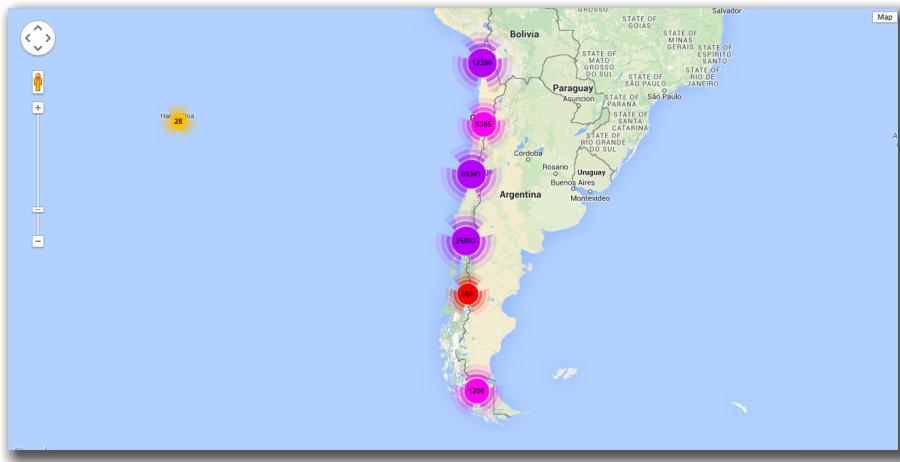


Figura 5.3: Mapa con los usuarios por ubicación geográfica

5.5.3. Captación de usuarios

El proceso de captación del conjunto de usuarios esta diseñado con la intención de obtener todas y todos los usuarios residentes en territorio Chileno dado que dicho catastro no existe en ninguna fuente oficial actualizada. El procedimiento se basa en las siguientes dos consideraciones:

1. Los generadores de información poseen una gran base de seguidores [47].
2. Es habitual que los seguidores de medios de prensa al querer difundir una información le escriban un tweet a algún medio de prensa o figura de autoridad, 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, las cuales son explicadas a continuación:

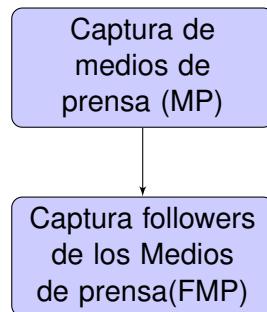


Figura 5.4: Diagrama conceptual con las etapas para la captación de usuarios.

5.5.3.1. Captura de los medios de prensa (MP)

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

- *ANP (Asociación Nacional de Prensa)* [18]: Asociación gremial constituida el 24 de agosto de 1951. Agrupa a los principales diarios y revistas del país.
- *ANARCICH (Asociación nacional de Radios Comunitarias y Ciudadanas de Chile)* [22]: Es el organismo que agrupa a 300 radios comunitarias y ciudadanas de todo el país.
- *ARCHI (Asociación de Radiodifusores de Chile)* [20]: 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, ANARCICH y ARCHI) cuentan con directorios públicos [19] [17] [21] 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
  
```

Tras aplicar este algoritmo humano, los resultados obtenidos fueron los siguientes:

Asociación	Nº Miembros con cuentas	Nº miembros totales
ANP	43	0,00 %
ANARCICH	23	2,09 %
ARCHI	XX	XX

Cuadro 5.4: Cantidad de miembros de las distintas asociaciones de prensa en Chile con cuentas en Twitter al 5 de junio del 2014.

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

Tras obtener la lista de medios de prensa, mediante la API de Twitter se recopilan los seguidores de cada uno de los medios de prensa. El algoritmo continuación realiza esta tarea requiere de realizar pero debe incluir intervalos de pausas para respetar las restricciones de número de solicitudes por hora que impone la API.

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 con más de 1,5 millones de seguidores la petición a la API 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 API⁵. Esta modificación fue realizada en base a las recomendaciones y comentarios disponibles en los grupos de desarrolladores de la librería [58] [11].

⁵ Al día 26/10/2015 no se encontraba disponible la mejora en el *Github* oficial de la librería [11]

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

5.5.4. Captura de tweets

El proceso de captura de tweets se realiza obteniendo los 100 últimos tweets de cada uno de los usuarios y usuarias recolectadas en la fase anterior, sin discriminación ni priorización alguna, 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 y posibles restricciones sobre si la cuenta no existe o es privada.

El algoritmo se demora en promedio 0,9658139 segundos en descargar los tweets de un usuario y almacenarlos en el sistema.

5.5.5. Procesamiento de los tweets

El procesamiento de los tweets se define en tres procesos:

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

5.5.5.1. Definición del tópico

Un tópico son las palabras claves que definen una búsqueda temática realizada por el administrador del 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.

5.5.5.2. 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 contengan las palabras claves que definen al tópico. Posteriormente se eliminan los tweets que poseen una similitud 0.85 % en sus textos utilizando la librería *difflib*. Esta consulta se demora 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

```

5.5.5.3. Depuración de conjunto de tweets relacionados al tópico

El proceso de depuración del conjunto de tweets obtenidos en el proceso anterior considera dos etapas distintas dependiendo del volumen de los tweets involucrados.

Un gran desafío en la depuración de los tweets, es discernir si un tweet en particular se relaciona o no con la temática del tópico, o por el contrario, si responde a una temática completamente distinta (y fue relacionado con el tópico únicamente por concordancia textual). Por ejemplo, si el tópico se refiere a "la paralización del registro civil", con la keyword 'paro' se busca excluir todos los tweets que se refieran a un "paro cardíaco." de "la acción de levantarse".

Para aplicar un filtro efectivo se consideran las siguientes premisas:

- El administrador del sistema posee poco tiempo (una tarea como eliminar filtros es un trabajo repetitivo y monótono, que requiere varias horas hombres para su desarrollo).
- 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 y su análisis manual no demanda tiempo excesivo.

Para conjuntos de 300 tweets o más se considera que su depuración manual es muy extensa y debe ser automatizada. Para su automatización se implementa un clasificador de Bayes-Naive. El clasificador Bayer-Naive utilizado es una implementación de la librería Python Textblob [70].

Los tamaños de los distintos conjuntos dependiendo del tamaño del conjunto de tweets es el siguiente:

Nº tweets	Nº Entrenamiento	Nº Validación
$N \leq 300$	Manual	Manual
$0 \leq N \leq 430$	$N * 0,2$	$N * 0,3$
$430 \leq N$	200	100

Cuadro 5.5: Cantidades conjuntos del clasificador utilizado

Estas cantidades fueron determinadas de manera empírica en base a recomendaciones recogidas en foros y experimentos propios realizados con la librería de tal manera de obtener una precisión de clasificación siempre superior a 85 %.

Algoritmo 9 Clasificador Bayes-Naive para determinar si es miembro o no del tópico.

```

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);

```

La clasificación de los distintos tweets en las categorías *pertenece al tópico* y *no pertenece al tópico* se realiza mediante la función *prob_classify(tweet).max()* que retorna la etiqueta de la clasificación que posee mayor probabilidad según el clasificador para el tweet en específico [59].

5.5.5.4. Orden geográfico

El orden geográfico dota al prototipo de un filtro que privilegia a las fuentes cercanas al lugar indicado como origen del hecho noticioso en desmedro de aquellas que se ubican a mayor distancia geográfica.

El orden geográfico se basa en la posición geográfica del autor de cada tweet analizado en la sección 5.5.2. Inicialmente se obtienen las comunas con menor distancia a la comuna central determinada como origen del tópico y se aumenta progresivamente la distancia hasta clasificar todos los tweets de autores geoposicionados, los tweets sin ubicación son desplazados al final del ranking.

Algoritmo 10 Orden Geográfico

```

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

5.5.5. 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 de un tweet para informar al usuario respecto al evento en cuestión.

Para su diseño se considera la cantidad de re-tweets con que cuenta el tweet basado en el razonamiento abordado en [79] 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 y en el estudio de la intención realizado en [84] referente a que si la intención del re-tweet es difundir (intención buscada por el prototipo, tweets que busquen informar y difundir sobre el tópico), es bastante probable que posea una gran cantidad de re-tweet, no así los tweets conversacionales .

En el diseño también se considera la fecha de emisión del tweet, implementando un sistema de tres clasificaciones basado en el ranking desarrollado en [23] con la diferencia que se utilizan tres grados de clasificación y no cinco.

El mecanismo de descenso 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 ← getMaxRT()
3:   minRT ← getMinRT()
4:   fechaMasReciente gets getFechaMasReciente()
5:   firstClass, secondClass, thirdClass gets dividirConjuntoPorRT(tweets)
6:   for tweet in firstClass do
7:     deltaFecha ← diff(fechaMasReciente, tweet.fecha)
8:     if deltaFecha ≥ 4 dias then
9:       tweet.clase ← tweet.clase - 1
10:    else if deltaFecha ≥ 10 dias then
11:      tweet.clase ← tweet.clase - 2
12:    for tweet in secondClass do
13:      deltaFecha ← diff(fechaMasReciente, tweet.fecha)
14:      if deltaFecha ≥ 4 dias then
15:        tweet.clase ← tweet.clase - 1

```

5.5.5.6. Panel de enlaces

Esta funcionalidad recoge todos los enlaces contenidos en los tweets del tópico con la intención de crear una sección de *bibliografía multimedia* de un tópico permitiendo acceder de manera directa 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 ← 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 ← 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 utiliza la librería lxml 5.2 de Python, que tras navegar

por la estructura del documento HTML se extrae el título, el 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 [26] y que el enlace no retorne error 404 (página no encontrada).

5.5.5.7. 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 la barra de búsqueda para ingresar las palabras claves del tópico, la fecha inicial del tópico y un menú para la ubicación del suceso.

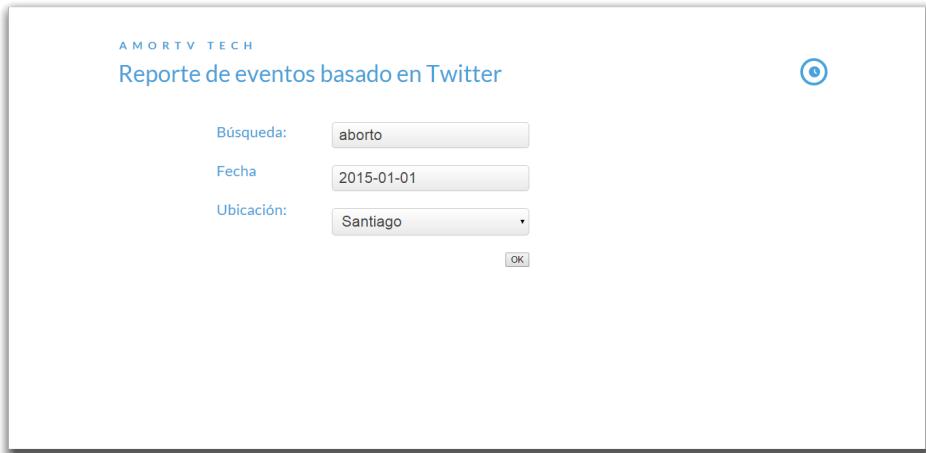


Figura 6.1: Vista nueva búsqueda tópico

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



Figura 6.2: Vista previa resultados de la búsqueda

- Vista de 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 timeline of tweets:

Fecha	Tweet
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/CVoRWNCWfo 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/lGDMsrc5fw 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.

AMORTV TECH

Reporte de eventos basado en Twitter

ABORTO [ORDEN RELEVANCIA]
PRENSA ON

Timestamp	User Handle	Tweet Content
2014-06-08 00:32	@ElardKoch	La ideología daña o mata la Ética Profesional eg. luego de aborto legal, se "persigue" a los objetores de conciencia
2014-10-27 15:24		Aborto en casos de violación sexual Sincerando el debate http://t.co/ZOIJ3FGjuv
2015-02-01 13:16	@GONZALOCOFLW	#EsteDomingoSoloSigoChilenos que se las juegan por los derechos reproductivos de las mujeres #Aborto3Causales
2015-01-31 19:04		@AbortoCeroEl Papa muestra su gran preocupación por el proyecto para despenalizar el #aborto en #Chile http://t.co/HFwFZ6UQXl
2015-01-12 17:29		Pobre Bither!! Ya con 2 abortos x algo terminara bien loca #AmorProhibido
2014-06-01 23:57	@margaritahantke	@ElardKoch el mito de la disminución del aborto con ley pro aborto. https://t.co/dmdnFdlhkmh

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.

AMORTV TECH

Reporte de eventos basado en Twitter

ABORTO [ORDEN RELEVANCIA]
PRENSA ON

Timestamp	User Handle	Tweet Content
2014-06-08 00:32	@ElardKoch	La ideología daña o mata la Ética Profesional eg. luego de aborto legal, se "persigue" a los objetores de conciencia
2014-10-27 15:24		Aborto en casos de violación sexual Sincerando el debate http://t.co/ZOIJ3FGjuv
2015-02-01 13:16	@GONZALOCOFLW	#EsteDomingoSoloSigoChilenos que se las juegan por los derechos reproductivos de las mujeres #Aborto3Causales
2015-01-31 19:04		@AbortoCeroEl Papa muestra su gran preocupación por el proyecto para despenalizar el #aborto en #Chile http://t.co/HFwFZ6UQXl
2015-01-12 17:29		Pobre Bither!! Ya con 2 abortos x algo terminara bien loca #AmorProhibido
2014-06-01 23:57	@margaritahantke	@ElardKoch el mito de la disminución del aborto con ley pro aborto. https://t.co/dmdnFdlhkmh

Figura 6.6: Vista de orden temporal

- Vista de 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.



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.



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.

Las cantidades de datos capturados son los siguientes:

Tipo de dato	Nº
Usuarios	650.000
Tweets	17.300.000

Cuadro 6.1: Cantidad de datos capturados

Características referentes a Twitter

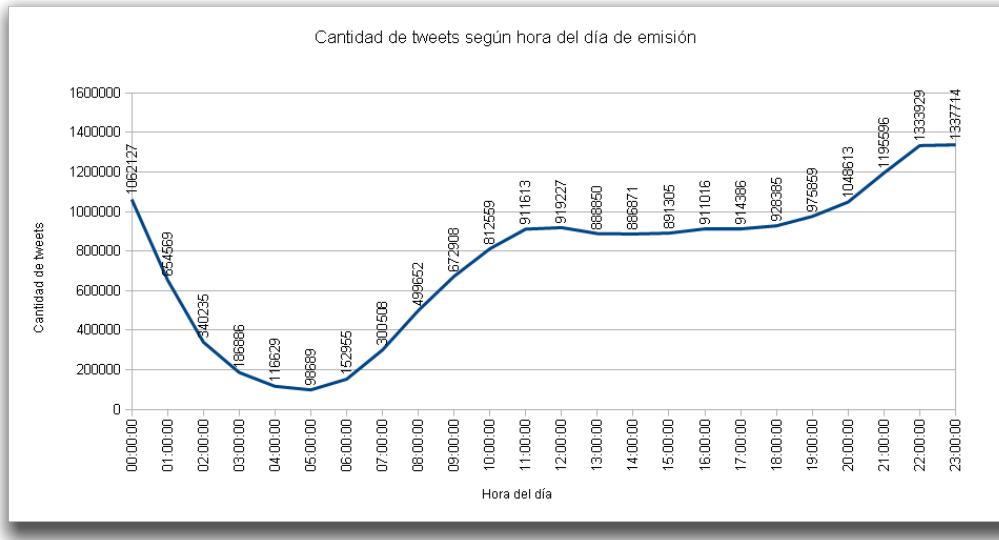


Figura 6.9: Distribución de tweets por hora de emisión

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 20:00 hrs. hasta las 23:00 hrs. mientras que el de menor emisión de tweets comprenden el periodo de horas desde las 3:00 hrs. hasta las 6:00 hrs. 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.

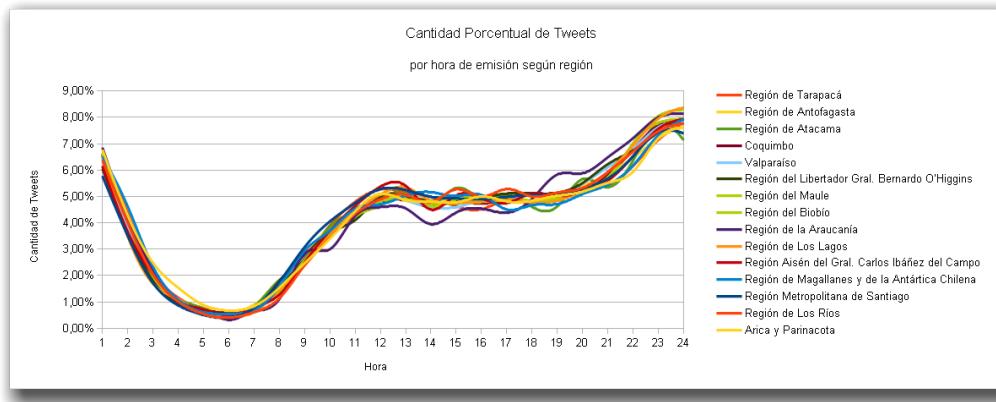


Figura 6.10: Distribución porcentual de tweets por hora de emisión según región

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 [79] 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 [79]

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 5.5.2 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 define anteriormente, se entiende por noticia "la comunicación de información seleccionada sobre un evento actual que posteriormente es presentado a través de cualquier medio de comunicación existente".[66] 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 [60] 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)

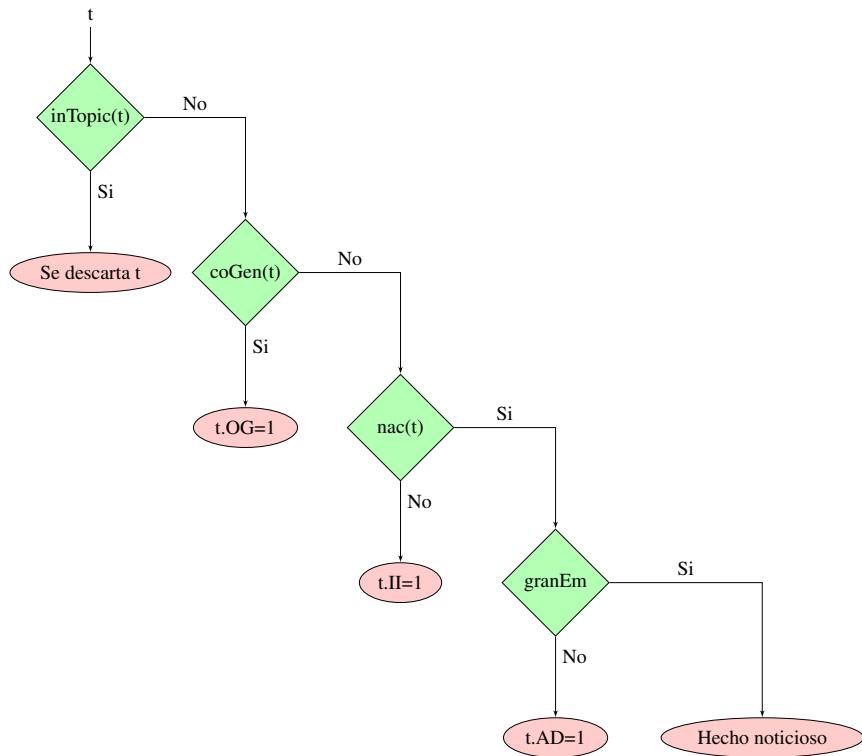


Figura 6.12: Procedimiento para determinar si un tweet se refiere a un hecho noticioso

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:

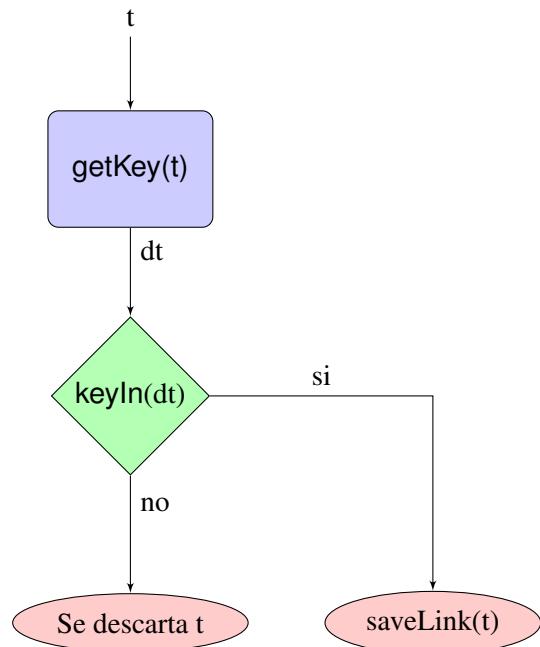


Figura 6.13: Procedimiento para verificar si el hecho noticioso haya generado una nota del medio modelo

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 Procedimiento 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;
```

El procedimiento anterior es ejecutado por un ser humano debido a la dificultad existente con las fuentes de verificación. Las únicas fuentes que poseen los medios modelo de notas históricas de acceso público, son sus respectivas secciones de búsqueda disponibles en sus sitios web [43] [10] [12].

Para analizar la información provenientes de estas fuentes se realizaron algoritmos de *scraping* directamente sobre los sitios web.

6.3.2. Caso 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 la clasificación manual realizados son los siguientes:

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

Cuadro 6.11: Clasificación referencia hecho noticioso

Con estos conjuntos de datos 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 aplica a la muestra el algoritmo explicado en 6.12. Con el cual se obtuvo la siguiente clasificación de los tweets de la muestra:

Clasificación	Total	Porcentual
Tweets que hacen referencia	108	34,95 %
Tweets que no hacen referencia	200	64,72 %
Total	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 %
Total	201	100 %

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

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. 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

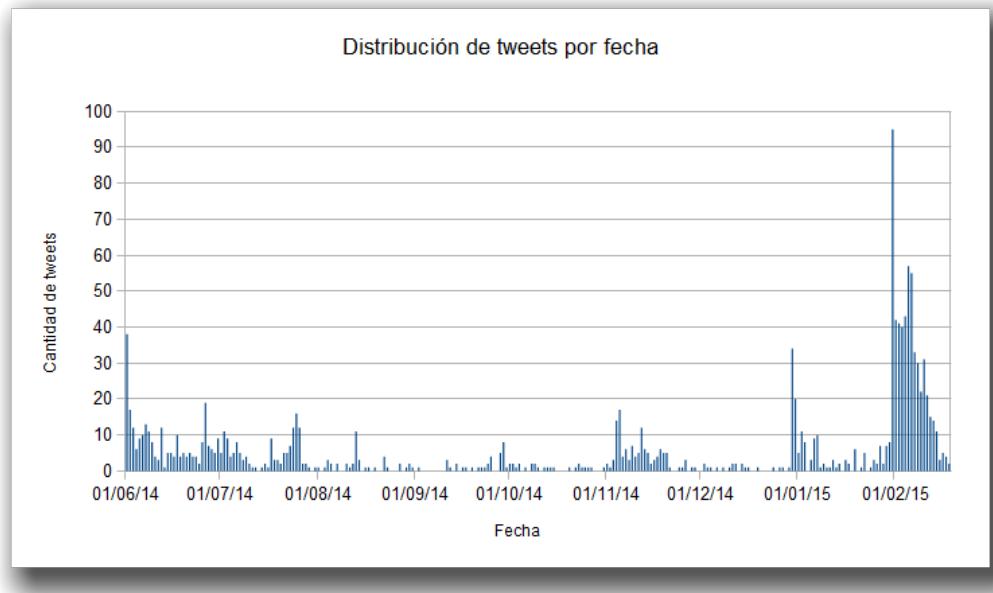


Figura 6.14: Distribución de cantidad de tweets por día

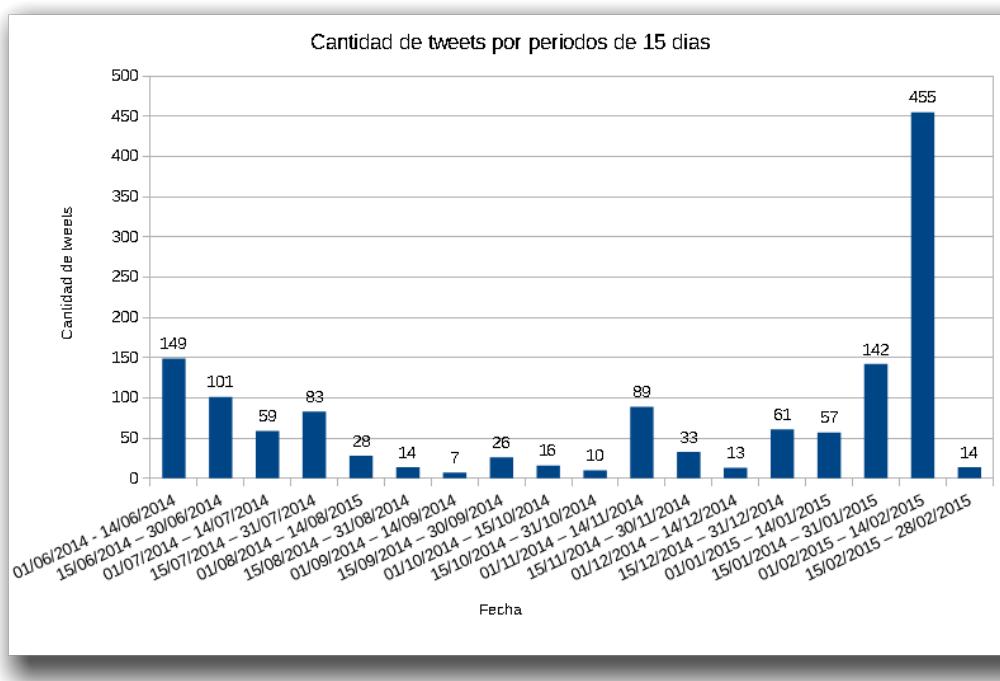


Figura 6.15: Distribución de cantidad de tweets por periodos de quince días

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

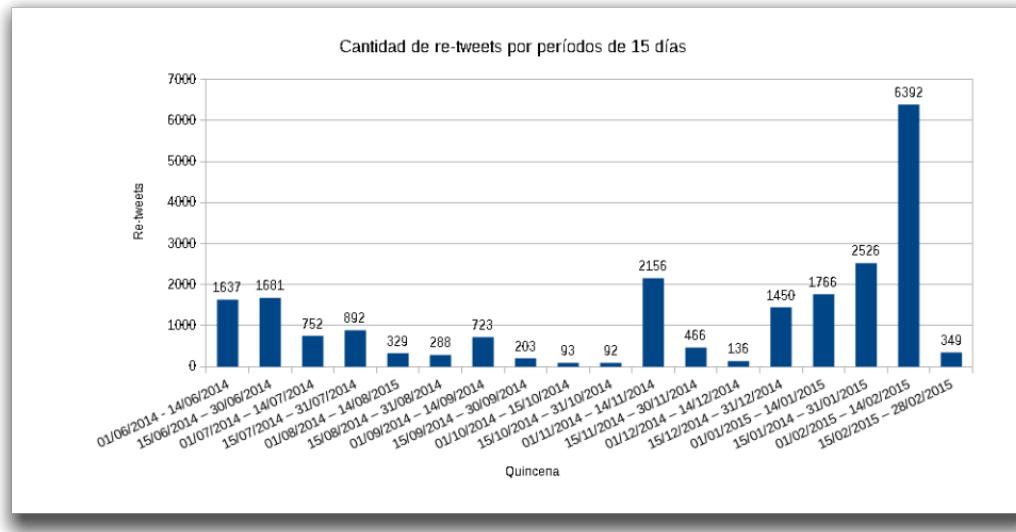


Figura 6.16: Cantidad de re-tweets por quincena

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 5.5.5.5. 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 “rascacé” 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://t.co/...
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/cP4ApM Zw33 #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 #TendráLibertad

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 (correspondiente al 26,8 %). De los cuales 20 se ubican en Antofagasta, 153 en Santiago. Referente a los tweets, los tweets geoposicionados corresponden a 382 de los 1382 tweets totales (correspondiente al 27,6 %).

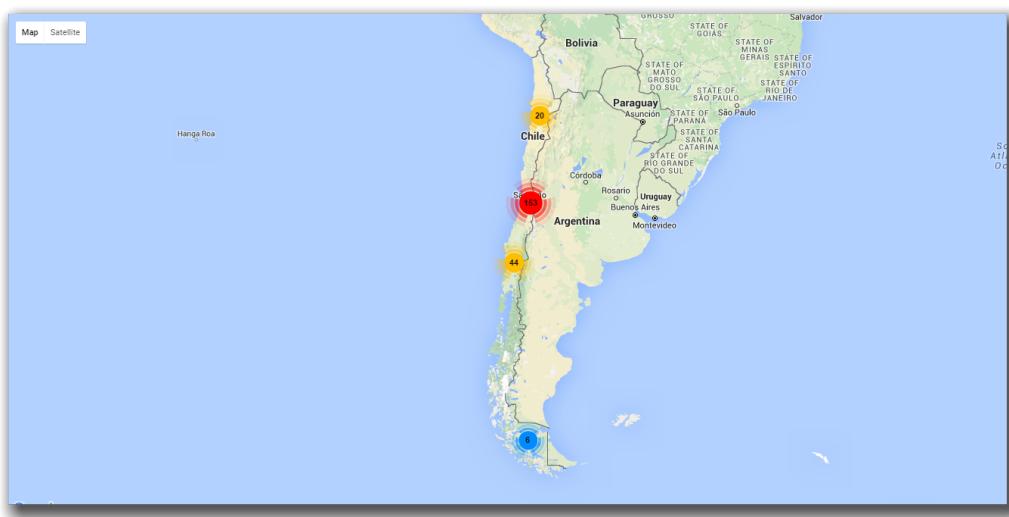


Figura 6.17: Distribución geográfica de los usuarios

Capítulo 7

Conclusiones

7.1. Conclusiones

7.1.1. Conclusiones técnicas

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, que se compone de una barra superior menú con 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 2 se aborda la discusión sobre de las razones de la existencia para los procesos de *gatekeeping*. Una parte importante de autores sostienen que 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, cada uno de los cuales fueron explicados en la sección .

En la sección 3 se analizan los distintos investigaciones y herramientas de fines similares

que tuvieran relación con aspectos abarcados en este trabajo. De manera general, es posible verificar a través de la creciente cantidad de estudios sobre Twitter el ascendente interés de la comunidad científica sobre esta red social. Muchos de los trabajos revisados abordan de distintas maneras la categorización de la información para dar solución a la problemática del valor de la información ante 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 permite 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.

Respecto 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 desarrollado mediante el ranking geográfico explicado en 5.5.5.4. El método implementado en este trabajo logra relacionar el 17,54 % de los usuarios totales con una provincia específica de Chile mientras que para el caso de prueba los tweets con localización corresponden al 26 % de los tweets del tópico. Existen variadas técnicas utilizadas para mejorar este relacionamiento en los diversos estudios analizados [9] [53] pero debido a su complejidad se pretende abordar como desarrollo futuro.

Respecto a los resultados abordados en 6 se verifica que Twitter es una fuente rica en información para la elaboración de reportes de un hecho noticioso (desde la cual se pueden filtrar efectivamente los tweets no relacionados), no sólo limitada a las noticias redactadas por la prensa convencional desde su visión editorial sino además de un gran conjunto de opiniones e insumos como enlaces externos que permiten profundizar la información. En el caso de prueba se observa que el 47,57 % corresponden a opiniones generales sobre el tema, el 34,95 % de los tweets se refieren a una noticia, , el 11,9 % corresponden a aportes a la discusión y el 5 % corresponden a aportes internacionales.

El ejercicio de informarse sobre los hechos noticiosos es fundamental para la generación de opinión ciudadana, una herramienta como la desarrollada, contribuye a esta labor en cuanto facilita el acceso a informaciones difundidas por otras y otros ciudadanos y pone a sus disposición un reporte que no solo hace referencia a noticias cubiertas por los medios de prensa convencionales sino que también opiniones, puntos de vistas y enlaces de documentación complementarios.

7.1.2. Consideraciones y discusión sobre las conclusiones

En la sección 2 se profundiza en el interesante y activo debate sobre la influencia real en el usuario y en la construcción de una noticia de los procesos de *gatekeeping*. El grado del impacto en el usuario depende de múltiples variables, entre ellas el ejercicio propio de informarse de cada persona que se dispone a informarse sobre un suceso noticioso. Una visión interesante es la que plantea Ramonet en [61] referente a que el ejercicio de informarse seriamente requiere esfuerzo y es una ilusión conseguirlo de manera cómoda, como supone la televisión. Es preciso para aprehender toda la complejidad de un suceso recordar los datos fundamentales de un problema, sus antecedentes históricos y su trama social y cultural. Esa misma filosofía sobre informarse, es la que da forma al prototipo al reunir en un mismo espacio comentarios, visiones y opiniones de distintas usuarias y usuarios ordenados y tratados con procesos de *gatekeeping* transparentes además una lista de enlaces externos donde profundizar o complementar puntos de vistas recogidos de múltiples y variadas fuentes. Estos mismos aspectos abren interrogantes sobre la fortaleza de la arquitectura del prototipo desarrollado ¿no es acaso una debilidad importante que solo cuente con una fuente de información como es Twitter? ¿no se expone a caso al *gatekeeping* proporcionado por Twitter?

Aún cuando Twitter, es una de las pocas redes sociales que trabaja constantemente en sus 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.

¹“Creemos 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” [76]

- **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[29], 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) [76].

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

Otro aspecto relevante considerado en este trabajo se refiere al origen del gatekeeping, si corresponde a un motivo operativo o ideológico, durante el desarrollo de este prototipo se verificó operativamente la fluctuación entre estos dos extremos, para conciliar este conflicto se considera que la única solución coherente a esta situación es transparentar los procesos de *gatekeeping* a los usuarios-consumidores de noticias, 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, similar a las prácticas y principios de trasparencia del proyectos *Open source*.

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

7.2. Trabajo futuro

El presente trabajo posee 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 las y los usuarios y su escasa tasa de llenado del campo *ubicación* en sus perfiles, se considera necesario desarrollar enfoques con mejores resultados considerando no sólo datos relativos a los usuarios sino al contenidos de los tweets (Como el abordado en [9] [37] [24] [36]). Un enfoque interesante de analizar es la identificación de hashtags locales y palabras locales, fortaleciéndolo con temáticas exclusivas y delimitadas a dicha zona, para identificar y relacionar tweets con esa zona geográfica específica.

En cuanto al desempeño del prototipo se pueden realizar importantes mejoras. Una de las opciones más atractivas respecto a la mejora de su arquitectura (relación rendimiento, escalabilidad y precio) es incorporar algunos de los Servicios Web de Amazon (AWS). AWS son el conjunto de servicios escalables tanto en costos como en capacidad que ofrece Amazon, permitiendo la automatización de procesos y el acceso a hardware de alto desempeño a un costo accesible (ver anexo 8.1). Los servicios específicos 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 y 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, eliminando las limitaciones de almacenamiento de datos. Su uso estaría destinado para almacenar los datos captados desde Twitter mediante la API y la realización de respaldos de seguridad.

- **Amazon Kinesis:** Proporciona un servicio de procesamiento de datos en tiempo real a streaming de gran cantidad de datos, capturarando continuamente y almacenando terabytes de datos por hora a partir de múltiples de fuentes de datos.

La incorporación de este servicio puede contribuir significativamente en la reducción de

tiempos que toman las distintas fases de procesamiento de los datos, abriendo también una increíble oportunidad de re-estructuración completa del proceso de captación de tweets, migrando la captación desde la API REST (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):** 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 se plantea como conclusión 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 público y la publicación del código en modalidad de software libre. 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 contraposición a los cambios inducidos por los medios de prensa en el último tiempo, en los cuales han incluido reportes de noticias ciudadanas pero solo 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.

Un aspecto interesante a explotar en esta misma línea es referente al 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.

Capítulo 8

Anexo

8.1. Cotización en Amazon

Instancia	Detalle	Costo
EC2	Linux c4.2xlarge, On demand con 1 adelanto parcial de reserva, 20 % utilización por mes. Transferencia de datos de entrada 10GB/Mes, Transferencia de datos de salida 3GB/Mes	US\$1243
EC2	Linux t2.small, On demand con 1 adelanto parcial de reserva, 100 % utilización por mes. Transferencia de datos de entrada 10GB/Mes, Transferencia de datos de salida 3GB/Mes	US\$228.48
RDS	db.r3.large, 100 GB, 200 IOPS, Mysql, 100 % utilización por mes	US\$239.89
Total		US\$1345

Cuadro 8.1: Costo mensual servicio Amazon (Precios al 18/11/2015). Elaboración propia (Disponible en <http://calculator.s3.amazonaws.com/index.html#r=IAD&key=calc-98B46645-C989-4AFA-976E-AFB1A7EB7A56>)

8.2. Cuentas en Twitter de los medios de prensa

Medio de prensa	Usuario	Nº amigos	Nº seguidores
40 Principales - Copiapó	@40ChileOficial	702	174K
40 Principales - Isla de Pascua			
40 Principales - Osorno			
40 Principales - Puerto Montt			
40 Principales - Rancagua			

40 Principales - San Antonio			
40 Principales - Talca			
40 Principales - Temuco			
40 Principales - Villarrica			
Agricultura - Los Angeles	@agriculturafm	2407	40,4K
Agrovision Fm			
Antofagasta Televisión	@antofagastatv	1107	18,4K
Autentica - Frutillar			
Autentica - Rengo			
Bravo	@RadioBravoFM	37	87
Carolina - Viña Del Mar	@RadioCarolina	15	358K
Chilena - Santiago	@RadioChilena	44	70
comunicaciones zona Fm			
Cooperativa - Ancud			
Cooperativa - Angol			
Cooperativa - Arauco			
Cooperativa - Calama			
Cooperativa - Caldera			
Cooperativa - Casablanca			
Cooperativa - Chillán			
Cooperativa - Constitución			
Cooperativa - Copiapó			
Cooperativa - Curicó			
Cooperativa - Iquique			
Cooperativa - Linares			
Cooperativa - Los Vilos			
Cooperativa - Mulchen			
Cooperativa - Ovalle			
Cooperativa - Puerto Aysén			

Cooperativa - San Antonio			
Cooperativa - San Felipe			
Cooperativa - Santiago			
Cooperativa - Talca			
Cooperativa - Temuco			
Cooperativa - Tocopilla			
Cooperativa - Valdivia			
Cooperativa - Vallenar			
Cordialissima			
Crystal			
Crystal - La Ligua			
Crystal - Quillota	@RadioCrystalQTA	181	160
Cumbre	RadioCumbreFM	103	162
Dalcahue	DigitalFMChile	1617	3846
Digital Fm - La Serena			
El Conquistador - Santiago	FMConquistador	193	31,4K
En Voz Alta			
Enamorada			
Entre Ríos	ENTRERIOSRADIO	113	46
Estación 106			
Eva Fm	radioevafm	1130	1263
Festival	radio_festival	1292	24,1k
Fm Contigo Tus Clasicos - Caldera			
Fm Dos - San Felipe	FMDOS	251	56,4k
Fm Okey - Los Andes	FMOK	1812	6730
Gaminides			
Hotel Cordillera			
Imagina - San Felipe			
Imagina - Santiago	imagina881	1285	4474

Imaginación Fm	imaginacionFM	276	182
Impacto - Calama			
Indomita			
Infinita - Los Angeles	infinitafm		
Infinita - Puerto Montt		307	10,2k
Lincoyan			
Los Confines			
Madero Fm	radiomaderofm	4480	8445
Mágica - Talca			
Manantial Fm - talagante	MANANTIALtgte	106	488
Mi Radio - Paillaco			
Nueva Fm Super Stereo			
Nueva Maule			
Nuevo Mundo	RNuevoMundo	4673	9247
Onda Fm			
Parinacota			
Pudahuel - Los Andes	RadioPudahuel	20,1k	25,2k
Radio Acogida			
Radio Aconcagua	aconcaguaradio	117	5839
Radio Aconcagua A.M y F.M			
Radio Actual Fm			
Radio Agricultura - Santiago	agriculturafm	2405	36,4k
Radio Almeyda Fm			
Radio Ambrosio			
Radio Ambrosio Linares	ambrosiofm	27	1055
Radio Amiga	radio_amiga	14	583
Radio Anahí	RADIOANAHICHILE	372	217
Radio Angel	RADIOANGEL3	90	79
radio Antares			

Radio Apocalipsis	fmapocalipsis	960	754
Radio Araucana			
Radio Artesanía	ARTESANIAFM	526	329
Radio Atractiva Fm	radioattractiva	892	1320
Radio Austral			
Radio Ayer - Rio Negro	Radio_Ayer	499	263
Radio Azul			
Radio Bahia - Taltal	Bahia_Radio	50	160
Radio Balneario			
Radio Beethoven 96,5 F.M.	radiobeethoven	3804	3888
Radio Bravissima F.M. Digital Stereo			
Radio Buena Nueva	RadioBuenaNueva	232	2412
Radio Buena Onda	radiobuenoonda	35	68
Radio Calama			
Radio Camila - Limache			
Radio Camila - Los Angeles	983camila	2	139
Radio Canal 95	canal95	376	2862
Radio Candelaria			
Radio Canelo 149 Am.			
Radio Cappissima	cappissima	195	923
Radio Caribe FM	caribefm	63	185
Radio Caricia			
Radio Carillon	radiocarillon	24	140
Radio Carnaval	Carnaval_FM	106	4034
Radio Carnaval - Ovalle			
Radio Carnaval - San Felipe	Carnaval_FM		
Radio Carolina - San Antonio			
Radio Carolina - Temuco			
Radio Carolina - Villarrica			

Radio Carolina Santiago	RadioCarolina	15	358K
Radio Centenario			
Radio Centinela	RadioCentinela		231
Radio Cobre Mar	cobremar_radio	303	185
Radio Colchagua			
Radio Comunicativa	radcomunicativa	1149	2313
Radio Condell	radiocondell	4482	10,1K
Radio Conifera			
Radio Contigo Fm	RadioContigoFm	2003	1240
Radio Copihue			
Radio Cultural			
Radio Del Lago			
Radio Desierto	desiertofm	1616	4366
Radio Difusora Dinámica Ltda.	radio_dinamica	780	1781
Radio Digital	DigitalFmChile	1618	3847
Radio Digital Fm - Diego de Almagro			
Radio Dinámica - Iquique			
Radio Duna - Viña Del Mar	RadioDuna	579	67,1K
Radio El Conquistador - Iquique			
Radio El Faro FM.			
Radio El Mundo			
Radio Ensenada Fm			
Radio Entrevalles			
Radio Esperanza - Temuco	esperanzafm	238	303
Radio Estrella Del Norte	estrelladelno	359	231
Radio Exodus			
Radio Fiessta	fiessta909	1197	5383
Radio Fm Mix			
Radio FM Plus			

Radio Fm Siete (Formato Español)			
Radio Futura F.M. Stereo	futurafmoficial	2230	3617
Radio Futuro - Punta Arenas	futurofm	710	148K
Radio Genesis - Andacollo			
Radio Genoveva 101.7			
Radio Gratissima	RadioGratissima	737	1157
Radio Guardia Marina Ernesto Riquelme			
Radio Horizonte	radiohorizonte	24,5K	76,1k
Radio Horizonte - Antofagasta			
Radio Horizonte - Temuco			
Radio Ignacio Serrano			
Radio Imagina	imagina881	1286	4618
Radio Integración			
Radio La Bruja FM			
Radio la Frontera			
Radio La Palabra			
Radio La Voz de La Costa	VozdelaCosta	1572	3333
Radio la Voz de la Tierra A.M.			
Radio Lanco FM.	radioLancofm	56	49
Radio Latina Fm			
Radio Lautaro	RadioLautaro	226	1380
Radio Libra			
Radio Lógika	LogikaFM	1836	2910
Radio Loncoche A.M.			
Radio Madrigal			
Radio Madrigal Fm			
Radio Malleco			
Radio Manía			
Radio Marcela			

Radio Maria	radiomariachile	284	1917
Radio Maxima - Antofagasta			
Radio Mirador	Radiomiradorfm	1875	1409
Radio Montecarlo - Iquique			
Radio Montecarlo - La Serena	montecarloc1	1100	3864
Radio Montecarlo - Ovalle			
Radio Montecarlo - Valdivia			
Radio Montecarlo - Vicuña			
Radio Monumental			
Radio Nacimiento 98,7 F.M.			
Radio Nahuelbuta	radionahuelbuta	234	1042
Radio Nativa	rad_nativa	33	242
Radio Nexo y Libra	radiolibraynexo	263	1716
Radio Norte Verde			
Radio Nueva Belén	nuevabelenfm	330	1368
Radio Nueve-Veinte			
Radio Nuevo Tiempo	ntchile	294	814
Radio Oceania			
Radio Pablo Neruda			
Radio Paloma	RADIOPALOMAFM	1311	27K
Radio Panamericana A.M.			
Radio Panorama			
Radio Parque Nacional A.M.			
Radio Paula Jaraquemada			
Radio Payne			
Radio Pewen FM	PEHUEN_FM	35	84
Radio Play FM - Antofagasta	play_fm	635	41,9K
Radio Play FM - Iquique			
Radio Play FM - La Serena			

Radio Play FM - Santiago			
Radio Portales Cb-118	RadioPortales	2302	8806
Radio Primordial			
Radio Principal Chuquicamata			
Radio Progreso			
Radio Pudahuel - Santiago	RadioPudahuel	22,1K	26,9K
Radio Puerta Norte			
Radio Pukara	RadioPukara981	144	108
Radio Raudal	raudalfm	562	137
Radio Renacer	Renacer1017	44	15
Radio Rio Claro			
Radio Rio Elqui			
Radio Rock & Pop - RANCAGUA	rockandpop	1119	84,2K
Radio Rock & Pop - SAN CLEMENTE			
Radio Romance			
Radio Romántica	Radio_Romantica	147	18,7K
Radio Romina			
Radio Sago & RadioSago	1131	11K	
Radio San Bartolomé	rsboficial	653	5289
Radio Santiago	radio_santiago	732	5293
Radio Sol	radiosolchile	1014	2243
Radio Somos Pichilemu	somospichilemu1	1161	1169
Radio Stellar Fm			
Radio Súper andina			
Radio Tiempo - Santiago	radiofmi tiempo	1184	19,4K
Radio Tiempo - Viña Del Mar			
Radio Topater	radiotopaterfm	476	141
Radio Trasandina			
Radio Trigal F M Stereo 103 9 Mhz	trigal_fm	249	1514

Radio Univ. de Antofagasta	radiouantof	714	473
Radio Universidad Austral			
Radio Universidad de Chile	uchileradio	2685	58,1k
Radio Universidad de la Frontera	UfroRadioTemuco	45	2452
Radio Universidad de Talca			
Radio Universitaria Fm	universitaria	1807	2108
Radio Universo - Santiago	RadioUniverso	144K	132K
Radio Ventisqueros - Chile Chico			
Radio Ventisqueros - Cochrane	Ventisqueros	91	1360
Radio Viaducto	RadioViaducto	406	948
Radio X - La Serena			
Radio X - Santiago			
Radio X - Viña Del Mar			
Radio xqa5			
Radio Zero S.A.	radiocero977	6561	61,8K
Radioactiva - Castro	RadioActivaFm	1019	82,5K
Radioactiva - Chillán			
Radioactiva - La Serena			
Radioactiva - Los Angeles			
Radioactiva - Osorno			
Radioactiva - Ovalle			
Radioactiva - Puerto Montt			
Radioactiva - Punta Arenas			
Radioactiva - Rancagua			
Radioactiva - San Antonio			
Radioactiva - San Felipe			
Radioactiva - Talca			
Radioactiva - Temuco			
Radioactiva - Valdivia			

RadioProyección	RadioProyecction	397	602
Radiovision	RadioVision997	14	23
Rinconada Fm			
San Sebastian			
Sociedad Radiodifusora Monterrey Ltda.			
Tiempo - San Antonio			
Tornagaleones	TomagaleonesFM	33	105
Ucv Televisión	ucv_tv	188	19,1K
Universidad de Santiago - Santiago	radiousach	3484	18,5K
Universidad de Tarapaca - Arica			
Universo - Iquique			
Universo - Osorno			
Universo - Temuco			
W Radio - Ancud			
W Radio - Arica			
W Radio - Chillán			
W Radio - Copiapó			
W Radio - Isla de Pascua			
W Radio - La Serena			
W Radio - Linares			
W Radio - Los Angeles			
W Radio - Osorno			
W Radio - Ovalle			
W Radio - Parral			
W Radio - Puerto Aysén			
W Radio - Puerto Montt			
W Radio - Punta Arenas			
W Radio - Rancagua			
W Radio - San Antonio			

W Radio - San Felipe			
W Radio - Talca			
W Radio - Tierra Amarilla			
W Radio - Valdivia			
W Radio - Villarrica			
W Radio - Viña Del Mar			
Cooperativa	cooperativacl	507K	1,73M

Cuadro 8.2: Medios de prensa ANP y características de sus cuentas en Twitter al 5 de junio del 2015.

Medio de prensa	Usuario	Nº amigos	Nº seguidores
Cuarta Colina	cuartacolina	49	79
Radiovision	FmRadiovision	17	41
Radio Alas	Radioalas	1987	1444
Radio Vanguardia	fmvanguardia	1	52
Radio Creativa FM	RadioCreativaFM	11	180
Radio Puro Chile	radiopurochile	25	242
Radio Paloma	RADIOPALOMAFM	1308	26500
Radio Enlace Sur	Radioenlacesur	326	113
Radio Interferencia	interferenciamfm	211	308
Casa Blanca	casablancafmm	222	219
Radio Ritoque	RitoqueFM	642	5062
Radio Algarrobo	radioalgarrobo		
La voz de Cerro Navia	lavozcerronavia	16	15
Radio Eclipse	radioeclipsefm	54	201
Radio Mater Dei	radiomaterdei	10	9
Radio Primavera	rprimaverafm	451	495
Radio Lorenzo Arenas	rlaradio	82	109

Radio Licanten	Radio_Licanten	922	607
Radio Espacios FM	espaciosfm	143	144
Radio Nueva Dichato	NuevaDichato	128	735
Radio Paula	Radio_Paula	28	115
Radio Pelom FM	RadioPelom	120	65
Radio Chiloé	radiochiloe	3140	8199
Radio Agricultura	agriculturafm	2405	39400

Cuadro 8.3: Medios de prensa pertenecientes a ANARCHIC y características de sus cuentas en Twitter al 5 de junio del 2015.

Nombre del Medio	Screen Name	Siguiendo (Following)	Seguidores (Followers)
La estrella de Iquique	laestrellaiqq	462	17900
La estrella de Loa	estrella_loa	82	2266
La estrella de Antofagasta	estrella_antofa	150	8613
Mercurio de Antofagasta	mercurioafta	1751	21600
Diario Chañarcillo	chanarcillo	1964	6597
Diario Atacama	diarioatacama	65	8570
Diario el Día	eldia_cl	580	38100
El Ovallino	elovallino	267	5851
Diario El Observador	eo_enlinea	614	14100
La estrella de Quillota	laestrelladeqta	1847	6773
La estrella de Valparaiso	laestrellavalpo	10	23800
Lider de San Antonio	lidersanantonio	289	7775
Diario el Labrador	diariolabrador	264	1600
Lider de Melipilla	lidermelipilla	465	657
La nacion	nacioncl	4116	220000
La Tercera	latercera	224000	1110000
El Mercurio	Emol	503000	1030000

La cuarta	lacuarta	36800	533000
Diario Financiero	DFinanciero	192	82800
Diario La Hora	DiarioLaHora	10300	207000
El Pulso	pulso_tw	816	31600
La Segunda	La_Segunda	3395	392000
El Rancaguino	elrancaguino	489	18400
La prensa de curicó	laprensacurico	1	5306
Diario El Centro	diarioelcentro	2333	22300
Diario Herald	Herald_Diario	9	131
Cronica de Chillán	CronicaChillan	88	17800
La discusión	ladiscusioncl	1409	19100
El Sur	elsurcl	645	26700
Diario Concepción	DiarioConce	1152	14300
La Tribuna	latribunacl	643	3354
El Austral	AustralTemuco	2255	40500
El Austral de los Ríos	austral_losrios	12	11700
El Austral de Osorno	austral_osorno	1128	15400
La Estrella de Chiloé	estrellachiloe	94	8424
El Austral	ellanquihue	747	20500
Diario El Llanquihue	ddivisadero	1401	3495
Diario de Aysén	diariodeaysen	318	1461
El Magallanews	elmagallanews	1343	4944
Diario el Pingüino	pinguinodiario	2027	19900
La Prensa Austral	LaPrensAustral	129	10400
La estrategia	estrategiacl	8467	39600

Cuadro 8.4: Medios de prensa pertenecientes a ANARCHIC y características de sus cuentas en Twitter al 5 de junio del 2015.

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] Summify Blog. Summifytip: Modifying summary sources and saying thanks. <http://blog.summify.com/2012/01/17/summifytip-modifying-summary-sources-and-saying-thanks/>. Accessed: 2015-11-10.
- [5] Shayne Bowman and Chris Willis. *we media; media; citizens media; participatory media; blogging*. Colection ourmedia. The Media Center, American Press Institute, 2003.
- [6] W. Breed. *Social Control in the Newsroom: A Functional Analysis*. Bobbs-Merrill Reprint Series in the Social Sciences, S34. Bobbs-Merrill, 1955.
- [7] 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.
- [8] 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.

- [9] 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.
- [10] Biobio Chile. Buscador bio bio. <http://busca.biobiochile.cl/>. Accessed: 2015-10-24.
- [11] DeWitt Clinton, Mike Taylor, and many contributors. Github python-twitter. <https://github.com/bear/python-twitter>. Accessed: 2015-10-25.
- [12] Grupo Copesa. La tercera. <http://www.latercera.com/resultadoBusqueda.html?q=>. Accessed: 2015-10-24.
- [13] Oracle Corporation. Website of mysql. <https://www.mysql.com/>. Accessed: 2015-10-25.
- [14] Oracle Corporation. Website of mysql workbench. <https://www.mysql.com/products/workbench/>. Accessed: 2015-10-25.
- [15] 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.
- [16] 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.
- [17] Asociación Nacional de la Prensa. Asociado a anp. <http://anp.cl/quienes-somos/asociados/>. Accessed: 2015-10-25.
- [18] Asociación Nacional de la Prensa. Website de asociación nacional de prensa. <http://www.anp.cl/>. Accessed: 2015-10-25.
- [19] Asociación de Radiodifusores de Chile. Radios socias de archi. <http://www.archi.cl/radios-de-chile.html>. Accessed: 2015-10-25.

- [20] Asociación de Radiodifusores de Chile. Website de archi. <http://www.archi.cl/>. Accessed: 2015-10-25.
- [21] Asociación Nacional de radios comunitarias y ciudadanas de Chile. Asociados a anarcich a.g. [http://radioscomunitariaschile.cl/Oficial/Copied-RED-NACIONAL.php](http://radioscomunitariaschile.cl/Oficial/Copied-Red-Nacional-de-radios-Comunitarias.php) <http://radioscomunitariaschile.cl/Oficial/Copied-RED-REGIONAL.php>. Accessed: 2015-10-25.
- [22] Asociación Nacional de radios comunitarias y ciudadanas de Chile. Website de anarcich a.g. <http://radioscomunitariaschile.cl/>. Accessed: 2015-10-25.
- [23] 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.
- [24] Mark Dredze, Michael J. Paul, Shane Bergsma, and Hieu Tran. Carmen: A twitter geolocation system with applications to public health.
- [25] ElementTree. Elementtree overview. <http://effbot.org/zone/element-index.htm>. Accessed: 2015-10-25.
- [26] Facebook. The open graph protocol. <http://ogp.me/>. Accessed: 2015-10-31.
- [27] Flipboard. Website corporative of flipboard. <https://flipboard.com/>. Accessed: 2015-10-18.
- [28] Django Software Foundation. Website of django. <https://www.djangoproject.com/>. Accessed: 2015-10-25.
- [29] 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/>.

- [30] Python Software Foundation. Pypi - the python package index. <https://pypi.python.org/pypi>. Accessed: 2015-10-25.
- [31] 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.
- [32] Tobias Futterer, Peter A. Gloor, Tushar Malhotra, Harrison Mfula, Karsten Packmohr, and Stefan Schultheiss. Wikipulse - a news-portal based on wikipedia. 2013.
- [33] 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.
- [34] Geofeedia. Website corporative of geofeedia. <https://geofeedia.com/>. Accessed: 2015-10-18.
- [35] L. Gomis. *Teoria Del Periodismo: Como Se Forma el Presente*. Paidos Comunicacion. Ediciones Paidos Iberica, S.A., 1991.
- [36] 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.
- [37] 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.
- [38] Antti Haapala. Python-levenshtein. <https://pypi.python.org/pypi/python-Levenshtein/>. Accessed: 2015-10-25.
- [39] Nadja Hahn. What good is twitter? the value of social media to public service journalism. Lse research online documents on economics, London School of Economics and Political Science, LSE Library, 2013.
- [40] T. Harcup. *Journalism: Principles and Practice*. SAGE Publications, 2004.
- [41] Alfred Hermida. Twittering the news. *Journalism Practice*, 4(3):297–308, 2010.

- [42] Alfred Hermida. Twittering the news: The emergence of ambient journalism. *Journalism Practice*, 4(3):297–308, 2010.
- [43] 24 horas. Buscador. <http://search.24horas.cl/search/?q=>. Accessed: 2015-10-24.
- [44] 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.
- [45] Ted. J. Smith III. La mordedura del perro guardian. *revista FACETAS*, 1991.
- [46] Twitter Inc. Documentation twitter developers. <https://dev.twitter.com/overview/documentation>. Accessed: 2015-10-25.
- [47] 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.
- [48] Andrés Azócar Z. y Andrés Scherman T. Julián González U. Encuesta de caracterización de usuarios de twitter en chile. <http://www.prensafcl.udp.cl/usuariotwitter.html> <http://www.prensafcl.udp.cl/usuariostwitter.pdf>. Accessed: 2015-11-07.
- [49] Infographic Labs. Twitter 2012. <http://infographiclabs.com/news/twitter-2012/#more-722>. Accessed: 2015-11-07.
- [50] Harold D. Lasswell. The structure and function of communication in society. pages 215–228, 2007.
- [51] K. Lewin. *Field theory in social science: selected theoretical papers*. Social science paperbacks. Harper, 1951.
- [52] W. Lippmann. *Public Opinion*. Harcourt, Brace, 1922.
- [53] 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.
- [54] 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.
- [55] Paper.li. Website corporative of paper.li. <http://paper.li/>. Accessed: 2015-10-18.
- [56] 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.
- [57] 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.
- [58] 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.
- [59] NLTK Project. Source code for nltk.classify.naivebayes. http://www.nltk.org/_modules/nltk/classify/naivebayes.html. Accessed: 2015-11-10.
- [60] Puroperiodismo. Menciones en twitter a medios chilenos de noticias. http://www.puroperiodismo.cl/?page_id=24464. Accessed: 2015-11-07.
- [61] Ignacio Ramonet. Informarse fatiga. *Le Monde Diplomatique*, Francia, 2000.
- [62] Stephen D. Reese and Jane Ballinger. The roots of a sociology of news: remembering mr. gates and social control in the newsroom. 2007.
- [63] J.F. Revel. *El conocimiento inútil*. Colección al filo del tiempo. Planeta, 1990.
- [64] Ramón Salaverría. *Redacción periodística en Internet*. Eunsa, 2005.

- [65] 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.
- [66] Clay Shirky. *Here Comes Everybody*. Penguin Books, London, 2008.
- [67] Smmry. Website of smmry. <http://smmry.com/>. Accessed: 2015-11-10.
- [68] Jetbrains s.r.o. Website of pycharm. <https://www.jetbrains.com/pycharm/>. Accessed: 2015-10-25.
- [69] 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.
- [70] 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.
- [71] Summify. Website corporative of summify. <http://summify.com/>. Accessed: 2015-10-18.
- [72] Summify. What is summify? <https://vimeo.com/15436623>. Accessed: 2015-11-10.
- [73] 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.
- [74] The Tweeted Times. Website corporative of the tweeted times. <http://tweetedtimes.com/>. Accessed: 2015-10-18.

- [75] Inc. Twitter. Acerca de la empresa. <https://about.twitter.com/es/company>. Accessed: 2015-11-07.
- [76] Inc. Twitter. Tweets still must flow. <https://blog.twitter.com/2012/tweets-still-must-flow>. Accessed: 2015-11-07.
- [77] Inc. Twitter. Twitter for newsrooms and journalists. <https://media.twitter.com/best-practice/for-newsrooms-and-journalists>. Accessed: 2015-11-10.
- [78] Twitterfall. Website corporative of storyful. <http://storyful.com>. Accessed: 2015-10-18.
- [79] 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.
- [80] K. Wahl-Jorgensen and T. Hanitzsch. *The Handbook of Journalism Studies*. ICA Handbook Series. Taylor & Francis, 2008.
- [81] D. M. White. *The gate keeper: A case study in the selection of news*. *Journalism Quarterly*. Bobbs-Merrill, 1950.
- [82] Arturo Arriagada y Patricio Navia. *Intermedios, medios de comunicación y democracia en Chile*. Ediciones Universidad Diego Portales. Ediciones Universidad Diego Portales, 2013.
- [83] 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.
- [84] Yuto Yamaguchi, Tsubasa Takahashi, Toshiyuki Amagasa, and Hiroyuki Kitagawa. Tu-rank: 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.