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

CARACTERIZACIÓN DE PERFILES INFLUYENTES EN TWITTER DE ACUERDO A TÓPICOS DE OPINIÓN Y LA GENERACIÓN DE CONTENIDO INTERESANTE

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

FELIPE ANDRÉS VERA CID

PROFESOR GUÍA:
JUAN VELASQUÉZ SILVA

MIEMBROS DE LA COMISIÓN:
FRANCISCO MOLINA JARA
IGNACIO CALISTO LEIVA

SANTIAGO DE CHILE
ABRIL 2015

RESUMEN DE LA MEMORIA
PARA OPTAR AL TÍTULO DE:
Ingeniero Civil Industrial
POR: Felipe Andrés Vera Cid
FECHA: 13/04/2015
PROF. GUÍA: Juan Velasquéz Silva

Resumen

Durante los últimos años en Chile ha aumentado el uso de Internet, de smartphones y de las redes sociales. Entre todas las redes sociales cabe destacar Twitter, dada la visibilidad que tiene al ser una red más abierta que otras. En Chile, el uso de Twitter se concentra en dos tipos: informarse y opinar. La cantidad de opiniones que se registran en Twitter es de gran interés para distintos actores del país, entre los cuales se encuentran empresas que utilizan Twitter como una herramienta de comunicación con sus clientes, para resolver quejas y dudas y hasta para realizar campañas de marketing viral en la red. Dada la masificación de Twitter y la gran cantidad de usuarios, existe la necesidad de poder saber el nivel de influencia de los usuarios y así poder priorizarlos en la resolución de sus necesidad como también poder hacer más efectivas diversas campañas de marketing.

Hoy en día, existen diversos servicios que realizan este tipo de tareas, como Klout o BrandMetric. Sin embargo, estos modelos miden la influencia de los usuarios de diversas formas, pero ninguno intenta vaticinar a los usuarios que se volverán influyentes en un futuro próximo. El presente trabajo consiste en definir una influencia en Twitter para luego ver se proyectaría en el tiempo, tomando como hipótesis que *es posible medir la influencia de un usuario a partir de su generación de contenido interesante*, para lograrlo se definió la influencia en la red de Twitter como *la capacidad de generar contenido interesante que repercute en la red social*. Viendo los modelos existentes se escogió uno y se modificó levemente para poder obtener un puntaje de lo interesante del contenido generado por un perfil.

Dado este modelo se generaron rankings sobre la influencia de un usuario en Twitter, además de rankings en agrupaciones de tópicos asociadas a *política* y *deportes*. No se pudo segregar en una mayor cantidad de tópicos por diversos motivos, por lo cual no se consideró que el modelo haya cumplido su objetivo de generar rankings de influencia para distintos grupos de tópicos. Luego, se realizaron los análisis de la predictibilidad para la influencia modelada, llegando a la conclusión que el periodo de datos es muy corto para poder predecir las series temporales.

Aunque los resultados pueden parecer desalentadores, el trabajo realizado deja un camino abierto para realizar otros enfoques y trabajos que son explicados en el capítulo final de la memoria. Así, se espera que una buena segmentación y priorización de perfiles puede servir para mejorar la resolución de problemas, encontrar perfiles que serán influyentes en determinados tópicos y focalizar campañas de marketing utilizando perfiles que no sean de un alto costo.

Agradecimientos

Quisiera partir agradeciendo a mis amados padres, Héctor y Alicia, por ser los mejores. Nada de esto hubiera sido posible sin el esfuerzo, sin la preocupación y sin la alegría que me han brindado durante todos los años de mi vida. Gracias a ustedes soy quien soy.

A mis hermanos, Héctor, Javier y Paulina, que gracias a cada sonrisa y cada broma han hecho de esta vida más entretenida. Siempre estaré agradecido de que sean mis hermanos y siempre estarán en mi corazón.

A Pamela por llegar a mi vida. Gracias por llenar este tiempo de momentos alegres. Que los frutos de este trabaja sean para el goce de ambos.

A mis amigos de la vida, en especial a Nicolás y Rocío, gracias por estar ahí y entregar toda esa felicidad que comparten al mundo.

Finalmente, gracias a todo el WIC, en especial a la Salita Sur y a mis profesores Juan Velásquez y Pancho Molina. Gracias por todo el tiempo, todos los consejos y toda la paciencia en la realización de este trabajo. Sigan apoyando a los próximos que vengan y confiando en ellos.

Tabla de Contenido

Índice de tablas

Índice de figuras

Capítulo 1

Introducción

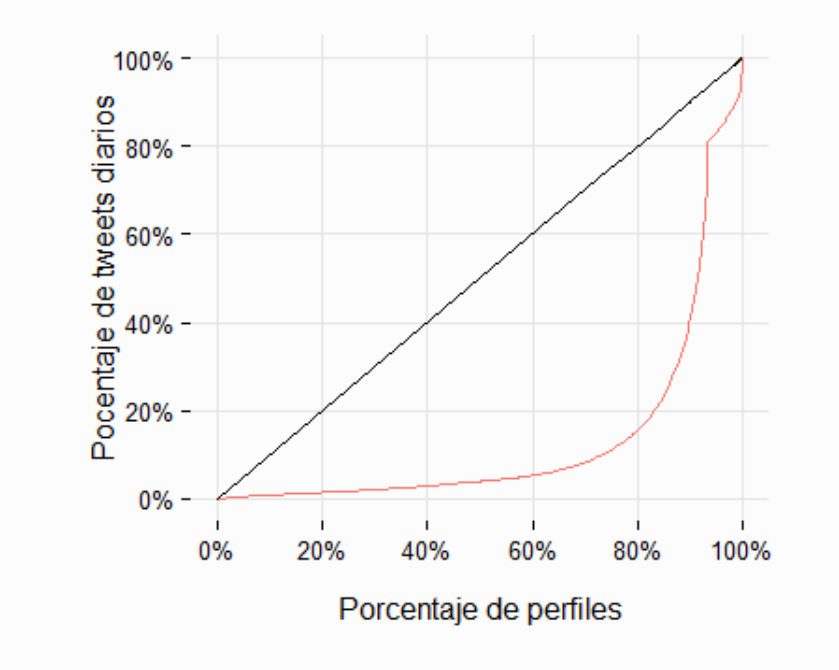
La adopción de tecnologías en Chile es envidiable para la región, nuestro país es actualmente el estado sudamericano con mayor penetración de Internet, siendo superior al 60 % para finales del 2014 [72][35]. Por lo otro lado, se tiene que la cantidad de celulares activos en Chile ya supera al número de habitantes [71]. Tantos celulares en la región son propiedad de cerca del 91 % de los chilenos, demostrando que muchas personas cuentan con más de un número activo, ya sea para el trabajo u otros asuntos personales. De hecho, el 78,2 % de los acceso a Internet se realiza por medio de un celular, esto gracias a la masificación del 3G y el 4G en el país[72]. Tanta conectividad dentro del país se ve reflejada en el día a día, con personas que pasan cada vez más tiempo comunicándose dentro de la llamada Web 2.0. Según un estudio del PEW Research Center, de los chilenos que usan Internet en su celular, un 76 % usa alguna red social[10]. Así mismo, dejando de lado los celulares, de las personas que usan internet en cualquier medio, 9 de cada 10 participan en alguna red social[13].

De las redes sociales utilizadas por los chilenos cabe destacar Twitter. En esta red de microblogging uno puede expresar sus opiniones en 140 carácteres y compartirlas con cualquier persona del mundo, lo que ha provocado que muchas celebridades y marcas usen este medio para potenciar sus imagenes y, a su vez, muchas personas usan este medio para poder compartir sus opiniones. El impacto que ha tenido esta red social ha provocado que diversos medios de comuniación estén al tanto de lo que ocurre a través de Twitter. Es común hoy en día ver en la prensa escrita o televisada el impacto en *las redes sociales* que generan diversas noticias de caracter nacional e internacional, mostrando desde *nubes de palabras* hasta las opiniones de diversos usuarios. La importancia que ha generado las redes sociales ha ido de la mano con su adopción en el país, aumentando año a año[13], llegando a casi dos millones de cuentas activas[1] en el país[41].

Una de las críticas que han surgido hacia Twitter es la poca participación de sus usuarios en la red social y, en efecto, en Chile se tiene que alrededor del 50 % usa la red social para poder informarse, mientras que en segundo lugar, con alrededor del 25 % de los usuarios, se usa Twitter para debatir y expresar opiniones[24]. En la figura 1.1 se aprecia como el 80 % de los usuarios no alcanzan a generar el 20 % de los tweets que se generan en un día, lo cual

[1]Donde activas son las cuentas detectadas por IBM para Chile en un lapso de 22 días para un estudio dado en el 2014.

Figura 1.1: Cantidad de tweets diarios que generan un grupo de usuarios en Twiter



Fuente: Elaboración propia a partir de los seguidores de las cuentas *tv_mauricio*, *latercera*, *biobio*, *elmostrador*, *canal13*, *tvn* y *mega*.

refuerza que existe un gran grupo de perfiles que solo consumen información.

Esto nos da a entender que esta red social es rica en opiniones de usuarios, y aunque no todos opinen, la mayoría usa principalmente Twitter para poder informarse, por lo que se cuenta con relaciones de influencia entre los usuarios donde unos pocos dan información a muchos otros. Esto se ve reflejado en la red en y su sistema de seguidores, donde pocos usuarios tienen una gran cantidad de seguidores y la mayoría de los usuarios sigue a una gran cantidad de perfiles, aunque tener muchos seguidores no es un predictor de influencia en la red[11].

1.1. Planteamiento del Problema y Motivación

Dado el creciente uso de las redes sociales, y de Twitter en Chile, se tiene que existe valiosa información que puede ser extraída. Dentro de esta información, cabe preguntarse cuál es más valiosa que otra, o en otras palabras, quién genera información más valiosa que otra. Para responder a esta pregunta existen diversas empresas (como BrandMetric[2] o Klout[3]) que generan ranking de personas de acuerdo a distintos criterios de influencia. Así también

[2]Empresa chilena que genera reportes a partir del contenido generado en redes sociales, donde uno de ellos es un *ranking* de influencia. `http://www.brandmetric.com`

[3]Empresa estadounidense que ordena a los usuarios de redes sociales de acuerdo a un *score* de influencia. `http://www.klout.com`

se han generado diversas investigaciones [84][70] que han indagando en la mejor manera para poder determinar a las personas más influyentes dentro de la red social.

Muchos de estos algoritmos intentan obtener los perfiles más relevantes de acuerdo a tópicos y a los niveles de influencia que se tienen frente a otros usuarios. Tales modelos han presentado mejoras a intentar ordenar información relevante (como el algoritmo de PageRank[4]) al poder adaptar sus modelos a páginas como Twitter y hacerles los arreglos correspondientes. Incluso, algunos investigadores[86] han propuesto que es más importante identificar el impacto del contenido generado más que le creador del contenido mismo, abriendo nuevas posibilidades de investigación dentro del área.

Entre un usuario y otro usuario es interesante preguntar quién será inflyuyente en el futuro. Esta memoria pretender poder vaticinar los usuarios que se volverán influyentes y, para lograrlo, se hará un algoritmo que sea capaz de obtener un score para poder medir la influencia de los usuarios, ver la tendencia en el tiempo y poder determinar algún patrón (con tal vez otras variables) que ayuden a determinar tales perfiles que tengan cierta probabilidad de ser influyentes en el futuro.

Las razones para realizar esto son diversas, como por ejemplo influenciar en la opinión pública[37] o en la adopción de nuevas tecnologías[65]. La difusión boca-a-boca puede alcanzar amplios rangos de población si nodos influyentes participan de la difusión, por lo que se vuelve de gran interés poder captar a estos nodos. Lo anterior es reforzado por el aumento del uso de las redes sociales en Chile y el interés de diversas empresas de poder saber a qué clientes priorizar en la resolución de problemas. Además al poder encontrar perfiles influyentes emergentes, se pueden focalizar campañas de marketing y utilizar perfiles que no sean los de más alto costo.

1.2. Hipótesis de Investigación

La hipótesis de investigación plantea que es posible medir la influencia de un usuario a partir de su generación de contenido interesante. Además, también se plantea que existe la factibilidad de poder dar una probabilidad de certeza de que un usuario será influyente en el futuro en base a una definición de influencia dada.

1.3. Objetivos

1.3.1. Objetivo General

Diseñar un algoritmo que pueda generar de manera cuantitativa un score de influencia de perfiles en twitter de acuerdo a tópicos de interés, para así obtener la caracterización de

[4]PageRank es un algoritmo que se aprovecha de la estructura de una red para asignar relevancia a ciertos nodos. Si un nodo es referenciado por muchos otros, se considera de mayor relevancia que el resto.

perfiles que tengan alta probabilidad de ser influyentes.

1.3.2. Objetivos Específicos

- Establecer el estado del arte de las técnicas relacionadas con índices de medición de influencia de perfiles en Twitter
- Definir un modelo a utilizar que permita obtener un indicador de medición del impacto de perfiles en Twitter
- En base al modelo elegido, crear un algoritmo que permita obtener un score de influencia de perfiles en Twitter de acuerdo a tópicos de interés y, si fuese posible y necesario, agregar velocidad de propagación
- Crear un modelo que entregue un grupo de perfiles en Twitter que tengan alta probabilidad de ser influyentes en un futuro próximo
- Evaluar y concluir en base a los resultados obtenidos

1.4. Metodología

La metodología propuesta se encuentra fuertemente relacionada con los objetivos específicos planteados anteriormente. A continuación se describe la metodología:

1. Estudio del estado del arte:
   Realización un análisis de la información bibliográfica existente, y así se levantó información útil acerca de distintos modelos existentes que evalúan la influencia en redes sociales.
2. Definición del modelo base:
   Dada la cantidad de modelos existen que definen la influencia en Twitter, revisar las fortalezas y debilidades de los distintos modelos para así definir el modelo base a utilizar.
3. Creación de algoritmo de influencia
   A partir del modelo elegido, realizar un *score* de influencia por distintos temas de opinión basado en la generación de contenido interesante que repercute en la red.
4. Creación de modelo de predicción
   Dado el algoritmo de influencia ya funcionando, realizar análisis de series de tiempo para ver la predictibilidad de la influencia futura.
5. Análisis de resultados y conclusiones
   Finalmente se observaron los resultados obtenidos y se concluye a partir de ellos.

1.5. Alcances

El foco de la memoria se encuentra en la identificación de futuros perfiles influyentes chilenos en Twitter, por lo que se tiene que la información a utilizar será los usuarios chilenos

de esta red. Como ya se ha mencionado, existen diversos algoritmos para poder encontrar un score de influencia, por lo que la memoria se basa en la generación de contenido interesante que repercute en la red social.

Además, se debe tener cuidado con diversos tópicos sensibles de la minería de datos, como lo son la orientación sexual, orientación política, orientación religiosa, problemas de salud y cualquier otro tópico que pueda ser causante de discriminación.

1.6. Contribuciones

Las contribuciones de este trabajo serán las siguientes:

- Obetener un marco conceptual basado en el estado del arte de los modelos actuales en la medición de influencia de perfiles en Twitter
- Realizar un modelo que permite medir la influencia de los perfiles en Twitter en base a la creación de contenido interesante que repercute en la red social
- Evaluar la posibilidad de predecir la influencia, concluyendo que se necesitan más elementos para intentar predecir
- Finalmente, sentar las bases para realizar algunos trabajos futuros que se describen en el último capítulo

1.7. Estructura del informe

La estructura de este informse se presenta de la siguiente manera:

El presentecapítulo trata de la introducción a la memoria, tomando en cuenta el contexto y las generalidades. Se plantean los objetivos de la memoria y la metodología a seguir

El segundo capítulo describe, de la manera más simple posible, el marco teórico investigado que es usado en este trabajo, además de diversas metodologías empresariales para poder medir la influencia.

El tercer capítulo trata de algunos modelos publicados que abarcan metodologías para medir la influencia en Twitter y revelar contenido interesante.

El cuarto capítulo describe las métricas definidas para la medición de influencia, la extración de datos en Twitter y como se realacionan ambos puntos para poder entregar un ranking de influencia por usuario y por temas.

El quinto capítulo indaga en las opciones de predicción que existen de series temporales y las dificultades presentes en el trabajo.

El sexto capítulo realiza una evaluación de los resultados obtenidos además de algunos

análisis de sensibilidad para ver otros resultados que se generan modificando ciertas variables de los modelos.

Finalmente, el capítulo séptimo trata del cumplimiento de los objetivos de la memoria, los problemas presentados durante el proceso y el trabajo futuro que queda propuesto a partir del trabajo realizado. Se termina con una conclusión general del trabajo realizado.

Capítulo 2

Marco Teórico

En el siguiente capítulo se presenta la base teórica del proyecto desarrollado. Para ello se dará una descripción general de la Web 2.0 y de Twitter, para luego seguir con los elmentos de la rama de Extracción de Información presentes en la memoria al igual que los de Modelamiento de Tópicos. Finalmente, se presentan miradas de influencia privativas que llevan a distintas maneras de medir esta en Twitter y los métodos predictivos que fueron usados en este trabajo.

2.1. Web 2.0

La Web 2.0 es un término popularizado en el 2004 con el cual se describe a las páginas que van un paso más adelante que las antiguas páginas estáticas en la web. Los sitios Web 2.0 suelen cumplir ciertas características, como lo son[55]:

- Tener un diseño centrado en el usuario: es decir un diseño que intenta en lo posible satisfacer las necesidades del usuario y le da cierta libertad de personalización.
- Participación de los usuarios: es decir cada usuario de la web aporta en la generación de información de un sitio.
- Colaboración de los usuarios: es decir que existe una comunidad de usuarios que colaboran entre sí para la generación de información, un caso emblemático es el de Wikipedia.
- Web como plataforma: es decir que las aplicaciones web ya no se cargan en computador de los usuarios, sino más bien en los servidores y los resultados con plasmados en un navegador.
- Contenido dinámico: es decir que los servicios que existen en la web son dinámicos y proactivos. En definitiva, el término de web 2.0 hace referencia al cambio de paradigma que existía antiguamente, donde los usuarios generaban poca información y solo recibían lo que proporcionaban las páginas web, a pasar a una red donde los usuarios son los principales motores y generadores de contenido, lo que ha llevado a la creación de distintas redes sociales que son fuentes de información, entre ellas Twitter.

2.2. Twitter

Twitter[1] es una red social de microblogging que cuenta con cualidades que lo hacen de interés para su estudio. Por ejemplo, que es una red social asimétrica al contar relaciones optativas, es decir las personas pueden elegir a quien seguir y no suelen escoger quien es quien los sigue. Además, es una red de microblogging con un límite establecido de 140 caracteres, disponible mundialmente, multiplataforma y además pública[2][78]. Todo ello hace a Twitter una red social donde las personas pueden informarse y debatir.

Entre algunos de sus elementos principales se encuentran:

- Followers: Son los seguidores de un usuario en particular en Twitter.
- Following: Son los perfiles a los que un usuario sigue en Twitter. Esto se usa para seguir las publicaciones de cierto usuario de interés.
- Tweets: Son las publicaciones en Twitter, estando limitados a 140 caracteres.
- ReTweets: Abreviado RT en la red social, indica cuando un perfil reenvía un Tweet de otro usuario a su audiencia.
- Reply\Mentions: Cuando un usuario menciona a otro para tener una conversación pública en base a un Tweet se llama *reply* (respuesta), cuando un usuario menciona a otro en una publicación propia se llama *mention* (mención), estos términos son usados con un símbolo arroba (@) antes del nombre del usuario a mencionar.
- Hashtag [82, 22]: Los Hashtag (del ingés *hash* almohadilla y *tag* etiquetar) son usados para mencionar explicitamente que se está hablando de un tema en particular. Para ello se utiliza la almohadilla (#) anteponiéndose a una palabra, así la etiqueta se puede identificar de una manera más rápida tanto para usuarios como para el sistema, como por ejemplo *#terremoto*, *#CASOPENTA*, *#justinbiever*, entre otros.

2.2.1. APIs de Twitter

Twitter cuenta con diversas APIs[3] para poder realizar consultas a su contenido, de estas APIs las de mayor interés para este trabajo son la REST API y la Streaming API.

La REST API de Twitter[79], está desarrollada por el tipo de arquitectura de desarrollo web REST[19] definida por Roy Fielding en el 2000. Esta arquitectura presenta una manera que puede ser usada por cualquier dispositivo que entienda HTTP, por lo cual es fácil de usar y ha tenido un amplio uso. Dado ello, la REST API de Twitter provee acceso a distintos datos de Twitter, cada una de estas solicitudes son llamadas peticiones y cuentan con un límite de 150 peticiones en ventanas de 15 minutos. El soporte para esta API se encuentra de manera extra oficial en diversos lenguages, siendo uno de estos JAVA con la libreria Twitter4j.

[1] www.twitter.com

[2] Esto es por defecto, un usuario puede cambiar el estado de su cuenta a privada

[3] La interfaz de programación de aplicaciones (API por su sigla en inglés) representa la capacidad de comunicación entre componentes de software. Esto viene dado por un conjunto de llamadas que se pueden realizar a ciertas librerías de un servicio para obtener información, generealmente, de capas inferiores a capas superiores.

Con esta libreria se puede obtener la siguiente información de un usuario de Twitter:

- Obtener la URL de la imagen de perfil
- Obtener la fecha de creación del usuario
- Obtener la descripción del usuario
- Obtener la cantidad de tweets marcados como favoritos
- Obtener la cantidad de followers que tiene el perfil
- Obtener la cantidad de followings que tiene el perfil
- Obtener la ID única que twitter asigna al usuario
- Obtener el lenguaje preferido que usa el usuario
- Obtener la cantidad de listas públicas en las que está el usuario
- Obtener la locación del usuario si la tiene descrita
- Obtener el nombre del usuario
- Obtener el color de fondo en el perfil
- Obtener la imagen de fondo en el perfil
- Obtener la URL del banner del perfil
- Obtener los colores del perfil
- Obtener el nombre de twitter que tiene el usuario
- Obtener el último tweet del usuario
- Obtener la cantidad de tweets que ha realizado el usuario
- Obtener la URL que tiene el usuario en su descripción
- Testear si el usuario ha cambiado el tema de su perfil
- Testear si el usuario es privado
- Testear si el usuario está verificado
- Testear si el usuario tiene la georeferencia de sus tweets activada
- Testear si el usuario es traductor

Y los siguientes datos de un tweet:

- Obtener la fecha de creación del tweet
- Obtener cuantas veces el tweet a sido marcado como favorito
- Obtener la geolocalización del tweet
- Obtener la ID del tweet
- Obtener el nombre/ID del usuario a quien se le hizo la respuesta si así fuese
- Obtener la ID del tweet al que este fue respuesta
- Obtener el idioma del tweet
- Obtener el lugar si ha sido adjuntado
- Obtener la cuenta de cuantas veces el tweet ha sido retweeteado
- Obtener el texto del tweet

- Obtener el usuario creado del tweet
- Testear si es un retweet
- Testear si ha sido retweeteado
- Testear si el tweet contiene algún link marcado como sensible
- Testear si el tweet es favorito

Todo esta información es entregada en cadenas de texto en formato JSON. El formato JSON es una notación para el intercambio de datos cuya finalidad es que sea fácil de leer para humanos y fácil de procesar para máquinas[36]. Fue primeramente especificado y popularizado por Douglas Crockford a principios del 2001 y se ha vuelto muy popular como entrega de información donde diversas APIs, como la de Twitter y la de Facebook, entregan la información en este formato. En la figura 2.1 se muestra un ejemplo de una parte de un texto en formato JSON entregado por la API de Twitter, se puede apreciar como cada elemeto del JSON es fácilmente identificable.

Figura 2.1: Parte de un JSON entregado por la API de Twitter


```
{
  "in_reply_to_status_id_str": null,
  "in_reply_to_status_id": null,
  "possibly_sensitive": false,
  "coordinates": null,
  "created_at": "Sun Nov 09 02:15:24 +0000 2014",
  "truncated": false,
  "in_reply_to_user_id_str": null,
  "source": "<a href=\"http://twitter.com/download/android\" rel
=\"nofollow\">Twitter for Android</a>",
  "retweet_count": 0,
  "retweeted": false,
  "geo": null,
  "in_reply_to_screen_name": null,
  "entities": {
    "urls": [
      {
        "display_url": "instagram.com/chester_gold_c…",
        "indices": [
          9,
          31
        ],
        "expanded_url": "http://instagram.com/chester_gold_cat",
        "url": "http://t.co/gpKVKYZrgE"
      }
    ],
    "hashtags": [
      {
        "indices": [
```


Fuente: Elaboración propia a partir de datos extraídos de Twitter.

Por su parte, la Streaming API[80] permite obtener información en tiempo real de lo que sucede en Twitter. Esta API cuenta con diversos filtros que se pueden aplicar para afinar la búsqueda. Como manera de prueba, esta API cuenta con el modo *firehouse*, el cual entrega una cierta cantidad de Tweets públicos de manera aleatoria. Esta muestra aleatorio no suele ser de interés y tiene un uso de prueba de la conexión con la API. Entre los filtros aplicables para afinar la búsqueda se encuentran:

- Idioma del Tweet: El idioma de un Tweet sigue la estructura de BCP 47[60] para los identificadores de idioma, aquí Twitter realiza una identificación del idioma de cada

tweet y gracias a ello se puede filtrar los que se quieren recibir, así, por ejemplo se puede conectar con $language = es$ para obtener solo los mensajes que el servicio identifica en español.

- Mensajes de un usuario en específico: Esta opción permite seguir el comportamiento en la red de un usuario en específico, es decir los tweets que el crea, que retweetea y las respuestas (*replies*) que realiza. Además, este parámentro captura las respuestas que recibe el usuario y los retweets que son realizados a través del botón ReTweet de los mensajes que creó.
- Palabras a seguir: Usar palabras claves para refinar la aparición de tweets es muy útil si se requiere realizar algún análisis en específico, por ejemplo buscar palabras que contengan la palabra *falabella*. Así, la API permite limitar la aparición de tweets a las palabras que uno selecciona, pudiéndose utilizar operadores lógicos O e Y para la inclusión de términos.
- Lugares: Finalmente, dentro del streaming público, es posible obtener tweets en un rango de coordenadas dadas. Los tweets han de estar geolocalizados para que Twitter sepa su procedencia y sea viable utilizar este filtro. Se pueden concatenar más de los coordenadas, obteniendo la unión de ambos conjuntos de tweets a recolectar, no así la intersección.

Cabe destacar que no es posible utilizar los diversos filtros para obtener la intersección que se generaría de ellos. Si se utiliza más de un filtro en el streaming público se obtiene la unión de ambos conjuntos. Por lo tanto, si fuese necesario utilizar más de un filtro a la vez de debe obtener los mensajes de un primer filtro y luego aplicar un segundo para obtener la intersección deseada.

2.3. Extracción de información

La extracción de información (en inglés *Information Retrieval*) es una rama de las ciencias de la computación que se centra en la extracción y el almacenamiento de documentos de manera automática [20]. Hoy en día, la extracción de información ha variado a como se ha estado haciendo hace años gracias a la expansión de la internet [3]. Este amplio campo de la investigación no es el objetivo central de esta memoria, por lo cual solo se mencionarán diversos elementos que son de interés para el presente trabajo.

2.3.1. Crawling

Crawling es el proceso de recopilar datos desde la Web [39]. Un crawler, otras veces llamada *araña web*, es un software que se dedica principalmente a actualizar su propio contenido que tiene de ciertas webs o para indexar el contenido que tiene de algunas páginas. Muchos buscadores realizan este proceso para poder indexar las páginas y poder buscar dentro de ellas rápidamente [31].

2.3.2. Pre procesamiento de datos

La cantidad de datos que se pueden obtener de la Web hace que sea necesario una limpieza de los datos antes de pasar al procesamiento de ellos. Esta tarea, en text mining, se realiza en distintos pasos de acuerdo a los requerimientos existentes. Entre los procesos más usados se tiene:

Tokenización

La tokenización [26] es un proceso en el cual se divide el texto de entrada. Si se toma un documento como una cadena de caracteres, esta cadena se puede dividir en piezas del texto (usualmente una palabra) que es llamada *token*. La manera de definir este fragmento de texto es variada, pero usualmente se divide por palabras para luego ser procesadas. Para poder quedarse solo con palabras se suele denotar el espacio como un delimitador de palabras además de eliminar signos de puntuación que no son parte de un carácter. Así, el texto *"Un gato corría por un ratón."* quedaría dividido en los tokens *"Un"*, *"gato"*, *"corría"*, *"por"*, *"un"* y *"ratón"*.

Stemming

Stemming, del inglés stem, es un proceso en el cual se lleva a las palabras a la raíz de estas. Así, por ejemplo, *bibliotecario* y *biblioteca* luego de pasar por este proceso se transforman en *bibliotec* y ello permite trabajar con menos palabras para la clusterización de estas.

Uno de los algoritmos más usados es el de Porter [61] de 1980, este algoritmo propone eliminar sufijos de las palabras basándose en la gramática y en reglas de reemplazo. Aunque antes se habían desarrollado distintos algoritmos de stemming, el algoritmo de Porter presenta una manera simple y algo más efectiva que otros modelos más complicados, tal como se muestra en su publicación[61].

Porter representó las palabras en consonantes "c" y vocales "v". Además cada grupo de vocales vvv... es denotada como V, de la misma manera un grupo de consonantes mayores a 0 es denotado C. Gracias a ello se tiene que cualquier palabra se puede representar como:

* $CVCV...C$
* $CVCV...V$
* $VCVC...C$
* $VCVC...V$

Si se denota los paréntesis de corchete como una presencia arbitraria de consonantes o vocales, estas cuatro formas se pueden representar como:

$$[C]\,VCVC...\,[V]$$

Así, Porter usa $(VC)^m$ para denotar que VC se repite m veces en una palabra, llamando al valor m *medida*. Cada palabra tiene su *medida*, así $ARMADO$ que tiene *medida* 2 está formado por $VCVCV$ mientras que AMA está formado por VCV por lo que su *medida* es 1. Entre otros ejemplos se tiene:

$m = 0$	QUE (CV), Y (V), ME (CV), TE (CV), PROA (CV)
$m = 1$	AMA (VCV), ROMPE (CVCV), PAN (CVC), NUNCA(CVCV)
$m = 2$	ARMADO (VCVCV), PRIVADO (CVCVCV), PROBLEMA (CVCVCV)

Porter también plantea que cada remoción de un sufijo está dada por la regla:

$$(condición)\ S_1 \rightarrow S_2$$

Lo que significa que si una palabra termina con el sufijo S_1, y la raíz antes de S_1 satisface la *condición*, S_1 es reemplazado por S_2. La *condición* está dada usualmente en términos de la *medida*. Por ejemplo, en inglés:

$$(m > 1)\ EMENT \rightarrow$$

Donde S_1 es $EMENT$ y S_2 es nulo, por lo que si se tiene la palabra REPLACEMENT, esta se cambia por REPLAC, dado que tiene una medida de 2.

Porter propone las siguientes condiciones para la gramática inglesa:

- $*S$: La raíz termina en s
- $*v*$: La raíz contiene una vocal
- *d : La raíz termina en doble consonante
- $*o$: La raíz termina en cvc, donde la segunda c no es W, X o Y

El mismo Porter llevó este removedor de raíces a un lenguaje llamado Snowball, implementándolo en diversos lenguajes como Java o C y distintos idiomas como Español o Fránces[4].

Lematización

Aunque se suele utilizar el término de *stemming* y lematización como sinónimos, no lo son[44]. La lematización es un proceso bastante parecido al *stemming*, pero con un enfoque distinto.*Stemming* suele referirse a una heurística que quita el final de una palabra para poder agrupar aquellas que tienen la misma raíz, mientras que la lematización intenta llevar a la palabra a su lema utilizando el uso de un vocabulario y del análisis morfológico de las palabras, por ello se suele juntar dos palabras que podrían ser analizadas como una sola más que obtener aquellas con la misma raíz. Por ejemplo, en inglés, *better* podría tener el mismo lema que *good*; si se hubiera realizado stemming estás dos palabras no estarían relacionadas.

[4]El proyecto Snowball tiene diversos recursos en su página principal: http://snowball.tartarus.org/

Remoción de Stopwords

Las *stopwords* son palabras vacías sin significado útil (o muy poco) para el análisis de documentos[44, 3], estás palabras pueden ser como artículos, pronombres o preposiciones. En el pre procesamiento de texto estas palabras son usualmente eliminadas para poder quedar solo con los conceptos de lo que se está analizando. Cabe destacar no existen listas definitivas de stopwords y a veces no son eliminadas para realizar análisis por frases o para no tener problemas con algunas stopwords que presenten ambigüedad con algunos nombres, como en español *té* y *te*.

2.4. Modelos de Tópicos

Con la llegada de la digitalización de documentos aparecieron diversas organizaciones que pretendían guardan la información en formato digital. Una de ellas es JSTOR[5], la cual comenzó a indexar la información de diversas revistas científicas de manera digital. Ya con un largo catálogo de información surgió la necesidad, para los investigadores más modernos, de poder recorrer toda la información digitalizada de una manera más intuitiva y donde los documentos pudieran estar categorizados para así poder encontrar más información de intereses similares, o simplemente, para poder explorar documentos bajo el alero de un tópico en particular.

Ciertamente se podría etiquetar manualmente cada uno de los artículos de acuerdo al criterio de un grupo de personas, pero esto requeriría mucho tiempo solo para categorizar los documentos existentes y se tiene que cada año se generan numerosos artículos que requieren de un etiquetado. Con el fin de automatizar este proceso es que nacen los modelos de tópicos [7], que han servido para categorizar distintos tipos de documentos, desde publicaciones científicas a correos electrónicos; y también, cómo no, tweets [86].

Los modelos de tópicos usan ciertas definiciones en común, entre las que tenemos [7]:

- Palabra: Una palabra (o término) se define como una secuencia de letras de un alfabeto definido.
- Documento: Un documento se puede definir como una bolsa de palabras, donde ésta bolsa es un vector donde cada componente se asocia a una palabra del diccionario, mostrando la frecuencia cada término en el documento.
- Corpus: Un corpus es simplemente un set de documentos.
- Diccionario: Un diccionario es compuesto por todas las palabras que aparecen en un corpus.
- Tópico: Un tópico es definido como una distribución sobre un vocabulario fijo de términos. En otras palabras, un tópico se puede explicar como un conjunto de palabras relacionadas entre sí. Por ejemplo se puede tener las palabras $miau - pescado - gato$ con alta probabilidad para un tópico *gato*, mientras que las palabras $guau - hueso - perro$

[5] www.jstor.org

para otro tópico que se puede llamar *perro* con alta probabilidad. Ambos tópicos pueden tener todas las palabras, pero *hueso* se presentaría en el tópico *gato* con baja probabilidad.

A continuación, se tiene una pequeña descripción del modelo de tópicos Latent Dirichlet Allocation, desarrollado por Blei *et al.*.

2.4.1. Latent Dirichlet Allocation

Latent Dirichlet Allocation[8] (en adelante LDA) es un modelo generativo de tópicos de los más simples presentes hoy en día, por lo que su utilización es bastante amplia. Este modelo se diferencia de otros modelos precedores en que no se limita a asignar un tópico a cada documento[47][53], logrando con esta estructura adicional modelar de mejor manera lo que se ve frecuentemente, donde un documento puede estar escrito a partir de más de un tópico.

LDA asume que existen K tópicos que están asociados con un corpus, donde cada tópico contiene el diccionario completo del cual cada palabra tiene cierta probabilidad de pertenecer a un tópico. Como cada documento se forma por distintos palabras, finalmente se tiene que un documento se construye bajo una distribución de tópicos, mostrando que cada documento se genera por varios tópicos diferentes.

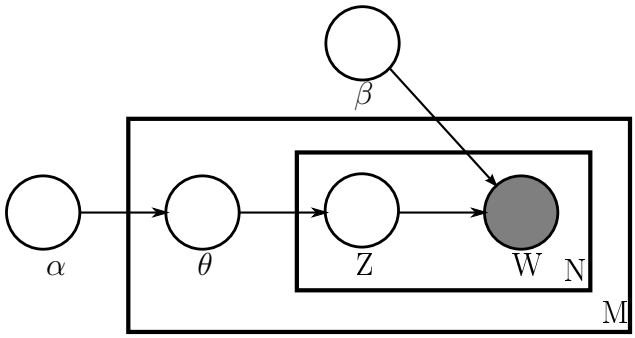
Para lograr lo anterior se asumen que existen *variables escondidas* y palabras que representan al documento en sí. Las palabras son la única variable observable del documento y todas las demás son generadas por el modelo para darle coherencia a la representación corpus presente.

LDA asume el siguiente proceso generativo para cada documento w en un corpus D :

1. Escoger $N \sim Poisson(\xi)$
2. Escoger $\theta \sim Dir(\alpha)$
3. Para cada una de las N palabras w_n:
   (a) Escoger un tópico $z_n \sim Multinomial(\theta)$
   (b) Escoger una palabra w_n de $p(w_n|z_n, \beta)$, una probabilidad multinomial condicionada al tópico z_n

Lo anterior se representa en la figura 2.2.

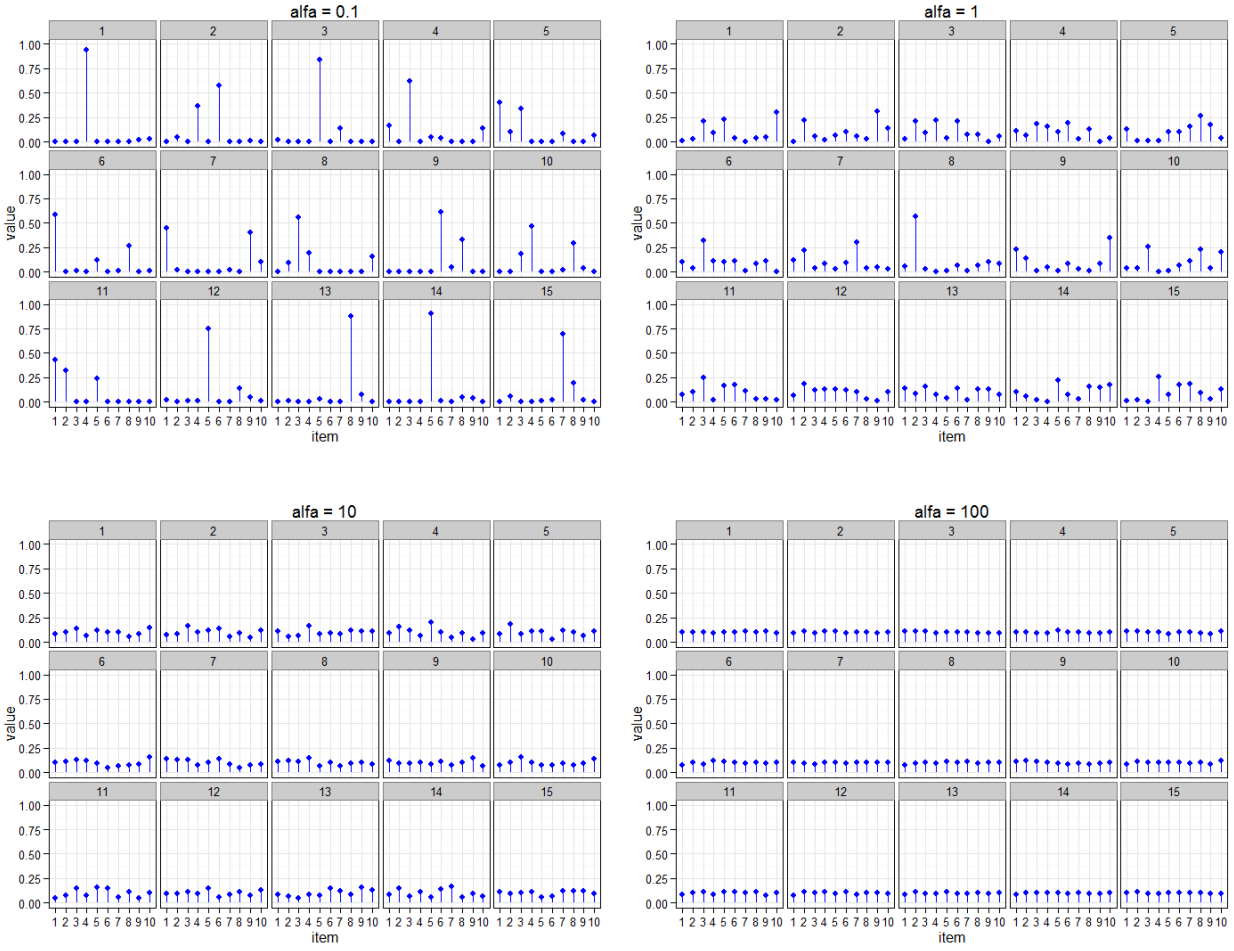
Figura 2.2: Representación gráfica del modelo LDA



Fuente: Elaboración propia a partir de [8]

Para poder ejecutar LDA es necesario incluir ciertos parámetros que no son determinados por el modelo, estos son los hiperparámetros de la distribución dirichlet α y β, como también el número de tópicos. Los hiperparámetros al ser vectores pueden contener distintos valores, pero se suele utilizar un valor simétrico para los hiperparámetros. Blei explica[7] que el valor de α y β se encuentran relacionados con la dispersión de las probabilidades de los documentos para cada tópico. En la figura 2.3 se pueden observar 15 simulaciones para 10 grupos con distintos valores de α. El eje x son las asignaciones a cada grupo mientras que el eje y corresponde a las probabilidades, es decir con el valor de α es posible controlar que tanto peso se le da a los grupos ya seleccionados cuando se quiere asociar un nuevo elemento a un grupo. En general con un valor alto de α un elemento es asociado a muchos grupos, mientras que con un valor pequeño se da el caso contrario.

Figura 2.3: Distribución Dirichlet con diversos valores de alfa



Fuente: Elaboración propia.

Griffiths y Steyvers[27] sugieren la utilización de un valor de $\alpha = 50/k$, donde k es el número de tópicos, y un valor de $\beta = 200/w$ o 0,1, donde w es el número de palabras en el vocabulario. Por su parte, Park y Ramamohanarao no hayaron una gran diferencia en los valores usados para α menor a 1[57], provocando que solo aumente considerablemente el tiempo de procesamiento. Diversas utilidades[45, 59] siguen la recomendación de Griffiths para seleccionar el valor de α, lo que es $\alpha = 50/k$.

Para el caso del número de tópicos, esto depende de lo que se requiera. Algunos autores utilizan un pequeño número de tópicos[89, 30], mientras que otros utilizan un largo número de tópicos[86, 27]. Es por ello que no es posible determinar un valor idóneo de tópicos, aun cuando existen acercamientos a una selección automática del número de tópicos[75], es necesario observar los resultados para ir afinando esta variable. Es posible que al elegir un número pequeño exista un sobrelapado de temas[27], mientras que en el caso opuesto una gran cantidad de tópicos puede fragmentar temas que se podrían considerar como uno solo. Es posible notar que el número de tópicos debe ser menor al número de palabras en el vocabulario, llegando al caso extremo donde cada tópico representa una palabra.

2.5. Influencia

Las definiciones de influencia varían de acuerdo al enfoque que se está mirando. George Mead ya en 1925[49] presentaba teorías del control social y argumentaba que el *control social* "dependerá del grado en que el individuo asuma las actitudes de aquellos que están implicados con él en sus actividades sociales", es por ello que desde hace décadas se ha querido desentrañar el quiénes son influyentes para poder afectar la opinión púbica[37].

En redes sociales se puede definir como la capacidad de afectar el comportamiento de otros[21], mientras que en difusión de información, influencia es la medida de popularizar información, es decir alguien influyente es quien propaga información en un rango más amplio de audiencia.

En la teoría de comunicación tradicional se establece que existe un grupo minoritario de personas, llamadas influyentes, que son sobresalientes en persuadir a los demás[65]. Esta teoría además predice que si uno es capaz de identificar a esos influyentes puede desencadenar grandes cascadas de información a un bajo costo de marketing[37]. Una mirada ya más moderna plantea que las personas de la era de la información realizan sus decisiones en base a las opiniones de sus pares y amigos[17] por lo que sería menos costoso identificar a las personas que influencian en pequeños grupos de gente.

Relacionados con la influencia, se dan distintos términos que son brevemente explicados a continuación.

2.5.1. Contenido Interesante

En social media se denomina contenido interesante a cuando la propagación de este no solo se realiza a los seguidores del autor, sino que a una larga audiencia. Un ejemplo de ellos es un mensaje de, por ejemplo, el cantante Justin Bieber, quien más que generar contenido interesante realiza comentarios hacia sus seguidores directos y ello genera *ruido* entre ellos, pero no pasará a otras comunidades.

2.5.2. Homofilia

La homofilia implica que el contacto entre dos personas con alguna similitud ocurre con mayor frecuencia que entre personas no tan símiles. Esta tendencia ha sido observada en decenas de estudios acerca de redes sociales. McPherson et al. [48] concluyó que la similitud entre distintos nodos alimenta las redes sociales de todo tipo. De igual manera, la evolución de la red social va muy ligada a la homofilia, donde las personas con mayor similitud se cohesionan más y las personas con menor similitud rompen sus lazos con mayor frecuencia.

En el caso de Twitter, la homofilia podría ser vista cuando un usuario sigue a otro por el interés que le provocan ciertos tópicos, y luego este le sigue de vuelta al ver que comparten intereses, esta aseveración es estudiada por [84] donde se llega a la conclusión que en Twitter

existe homofilia pero que no es imperante en la red social, o sea que existen individuos que de igual manera siguen a otros sin que exista una gran similitud entre ellos.

2.5.3. Mediciones privativas de influencia en Twitter

Desde la creación de Twitter han existido diversos medios que intentan darle determinar que usuarios tienen mayor influencia dentro de la red. Con el pasar de los años fueron surgiendo distintos servicios que priorizan a perfiles en Twitter. Entre los diversos servicios privados que existen se mencionarán tres importantes a nivel mundial, estos son Klout, Topsy y Kred; además de uno chileno, BrandMetric.

Klout

Klout[6] es una plataforma web que analiza diversas redes sociales para ordenar a las personas de acuerdo al nivel de influencia que tienen en las redes sociales. Por defecto, Klout indexa a todos los usuarios públicos de Twitter y les da un puntaje. Cuando uno ingresa a Klout puede vincular otras cuentas de diversas redes sociales que tenga, como Facebook[5] o Linkedin[46], con ello ayudando a la recopilación de información de la plataforma.

La manera en como Klout funciona es privada y ha sido críticada por diversos motivos, cómo el afán de mantener un score para las personas sin mayor certeza de donde proviene[1] o cómo ese mismo score puede aumentar considerablemente si uno agregar vincula muchas redes sociales a su cuenta[58]. Aun así Klout ha llamado gran atención de diversos medios y empresas, incluso llegando a promover la entrega de regalos y servicios por parte de distintas empresas a los usuarios que tengan cierto puntaje, estos regalos son llamados Klout Perks [63, 38]. Según la revista Wired, incluso se ha llegado al punto donde un VP de 15 años de experiencia fue rechazado por su bajo puntaje Klout [7].

Kred

Kred[8] es otra plataforma que mide el nivel de influencia del usuario, pero a diferencia de Klout, lo hace de manera transparente. Mide la influencia en base a las interacciones que uno tiene en la red social y con los tipos de cuentas con los que se interactúa, dando puntajes por estas acciones[40]. La tabla 2.1 muestra la asignación de puntajes por cada acción que se le da a un usuario, luego de obtener el puntaje total de todos los usuarios este se traduce al *score* de Kred que fluctúa de 1 a 1,000. Para mantener una distribución normal en el *score* asignado a los usuarios, se tienen escalones de puntajes para ir subiendo en el nivel de Kred. Así para subir de un *score* 800 a 900 se necesitan más de 5 veces el puntaje total obtenido hasta 800.

[6] https://klout.com/home
[7] What Your Klout Score Really Means URL: `http://www.wired.com/2012/04/ff_klout/all/`
[8] http://kred.com/

Tabla 2.1: Puntos ganados por acción en Kred

Interacción en Twitter	Puntos Ganados
Retweet o respuesta de un usuario con menos de 10,000 seguidores	10 puntos
Retweet o respuesta de un usuario con más de 10,000 seguidores	25 puntos
Retweet o respuesta de un usuario con más de 100,000 seguidores	50 puntos
Nuevo seguidor	1 punto

Fuente: Elaboración propia a partir de [40].

Kred separa a sus usuarios en *communities*, que son en el fondo temas de interés. Cada *communitie* tiene un usuario con un puntaje de 1,000. La manera en que los perfiles son asociados a un *communitie* se basa en las palabras que un usuario usa. Además si una cuenta en específico se sabe que es de un tema, se agrega manualmente a tal *communitie*.

Topsy

Topsy[9] es un servicio que mantiene un registro de todos los tweets que se han publicado en la historia. La herramienta tiene un uso de estadísticas y recuperación de información. Dado la gran cantidad de data existente, Topsy intenta mejorar sus resultados y análisis por el peso que tienen las personas en Twitter, es decir, por su influencia en la red.

Topsy mide la influencia como la probabilidad de que cada vez que alguien diga algo otro usuario tome atención a ello.[76]. Esta metodología no está explicitada en su página web, por lo que no es posible saber cómo ponderan los distintos factores que Topsy usa en su medición. Del mismo modo, Topsy no utiliza un ranking de influenciadores (a diferencia de Klout), sino que mantiene sus rankings de manera privada para poder darle más relevancia a los resultados de las búsquedas.

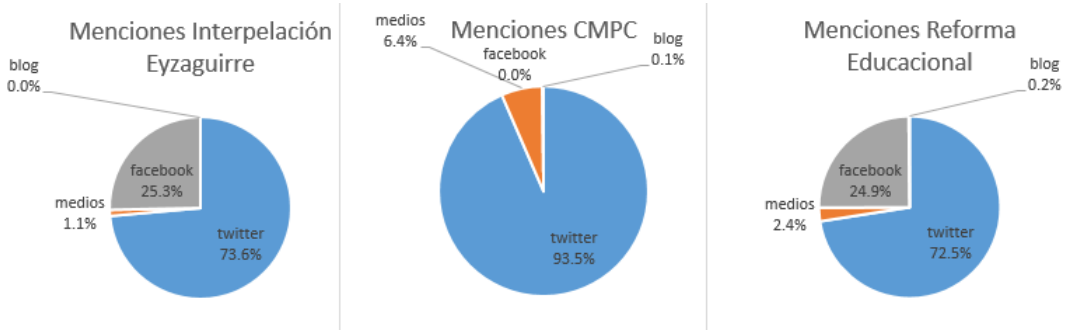
Brandmetric

Brandmetric[10] es una empresa chilena que entrega servicio de consultoría, reportes y alertas en redes sociales sobre marcas u otros temas. La mayoría de la información procesada por Brandmetric viene de Twitter[9], tal como se aprecia en la figura 2.4. Es posible notar que la gran cantidad de participación de Twitter se debe a la naturaleza más pública de sus datos, a diferencia de otros medios sociales como Facebook.

Brandmetric deja un apartado para medir a los más influenciadores de cierto tópico en Twitter. Para ello, calcula el porcentaje de las menciones y retweets de un usuario dentro del universo de todas las menciones: es decir, cuántas veces aparece @*usuario* en comparación a la suma de todas las otras menciones.

[9] http://topsy.com/
[10] http://www.brandmetric.com/

Figura 2.4: Porcentaje de Menciones para diversos temas en Brandmetric



Fuente: Elaboración propia a partir de datos extraídos de la demo de Brandmetric para Noviembre del 2014

2.6. Predicción de series de tiempo

La predicción plantea realizar análisis a una serie de datos para poder prever su comportamiento futuro, ayudando así a las desiciones y planificación[43]. Los métodos de predicción se pueden subdividir principalmente en tres: cuantitativos, cuando hay suficiente información disponible; cualitativos, cuando no existe tantos datos pero si hay una noción suficiente para la predicción; y finalmente, no-predecible, que se refiere a cuando existe poca información o no es viable. En la primera de estas categorías se encuentran las series de tiempo que son de interés para este trabajo.

Las series de tiempo son una secuencia de observaciones que estan ordenadas, como por ejemplo el PIB de un país, el precio de acciones o el clima de una región. Con estos datos se pretende identificar la estructura de dependecia temporal de las observaciones, y así poder utilizar eventos pasados para predecir eventos futuros o extraer el comportamiento cíclico de una serie de tiempo[28].

Para este enfoque de predicción se utilizan diversos métodos. Entre los más usados están los modelos autorregresivos integrados de media móvil (ARIMA, por sus siglas en inglés) y últimamente las redes neuronales artificiales (ANN, por sus siglas en inglés).

2.6.1. Modelo autorregresivo integrado de media móvil

Los modelos ARIMA son modelos estadísticos que utilizan variaciones y regresiones de datos estadísticos con el fin de encontrar patrones para una predicción hacia el futuro. Para poder formular un proceso de la forma ARIMA se deben cumplir dos condiciones necesarias en el proceso estocástico estacionario:

- El proceso no debe ser anticipante, es decir, que sus variables no dependan de las variables futuras.
- El proceso ha de ser invertible, es decir ,que la influencia de una variable pasada vaya disminuyendo con el tiempo

Además de estas condiciones, se tiene que un modelo ARIMA cuenta con diferentes partes que en su conjunto forman al modelo, la primera de ellas es la parte autorregresiva, donde un modelo autorregresivo de orden p se puede escribir como se muestra en la ecuación 2.6.1:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \mathrm{e}_t \tag{2.6.1}$$

donde c es una constante y e_t es ruido blanco. Este modelo recibe la notación de $AR(p)$. Un cambio de en los parámetros ϕ resulta en patrones diferentes de la serie.

Para las medias móviles se sigue un modelo que usa los errores predichos del pasado como se muestra en la ecuación 2.6.2:

$$y_t = c + \mathrm{e}_t + \theta_1 \mathrm{e}_{t-1} + \theta_2 \mathrm{e}_{t-2} + \cdots + \theta_q \mathrm{e}_{t-q} \tag{2.6.2}$$

donde e_t es ruido blanco. Este modelo tiene la notación de $MA(q)$, se puede notar que los valores de e_t no son observables.

Combinando ambos modelos se obtiene un modelo ARIMA no estacional, donde un modelo ARIMA (p, d, q) se puede representar como se muestra en la ecuación 2.6.3:

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \mathrm{e}_{t-1} + \cdots + \theta_q \mathrm{e}_{t-q} + \mathrm{e}_t, \tag{2.6.3}$$

en donde d corresponde a las diferencias que son necesarias para convertir la serie original en estacionaria, p es el orden de la parte autorregresiva del modelo, q es el orden de la parte de medias móviles del modelo[32].

Es así como el modelo ARIMA permite describir un valor como una función lineal de datos anteriores incluyendo un componente cíclio y estacional. Los precursores de estos modelos son Box y Jenkins, quienes recomendaban tener 50 datos como mínimo para realizar las predicciones[15].

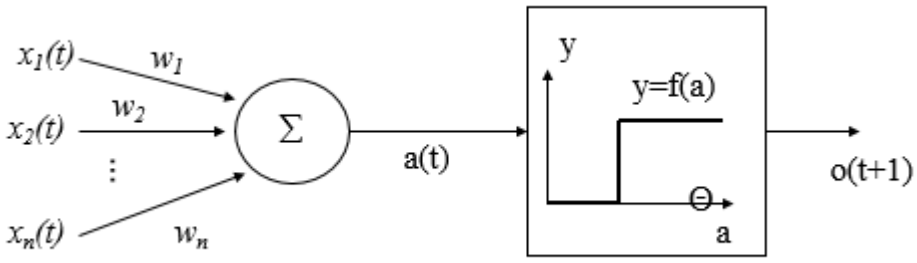
La metodlogía de Box y Jenkins se puede resumir en cuatro fases:

- La primera consta en identificar el posible modelo ARIMA
- La segunda consta de estimar los parámetros AR y MA en base al modelo seleccionado previamente usando máxima verosimilitud
- La tercera fase es de diagnóstico, donde se ve que los residuos sean un ruido blanco
- La cuarta fase es la predicción con el modelo escogido

2.6.2. Redes neuronales artificiales

Las redes neuronales artificiales vienen de la idea de imitar el funcionamiento de las redes neuronales biológicas, en este caso se trata de un sistema de interconexiones de neuronas

Figura 2.5: El modelo neuronal



Fuente: Elaboración propia a partir de [23].

que colaboran entre sí para producir un estímulo. Cada neurona recibe una cierta cantidad de entradas para luego emitir una salida. Este modelo se puede observar en la figura 2.5, donde se tiene un set de sinapsis donde cada neurona tiene un peso w que afecta en la señal recibida. Luego, hay un sumador que combina las señales de las sinapis de manera lineal y, finalmente, está la función de activación que limita el *output* de salida de la neurona, siendo generalmente en intervalos de 0 a 1 o de -1 a 1.

Las neuronas, por su parte, siguen procesos elementales que realizan una función definida. Estos procesos elementales se pueden organizar en capas que pueden ser de tres tipos.

- Capa de ingreso: Esta capa recibe la información y la lleva al siguiente nivel.
- Capas ocultas: Acá los procesos elementales procesan la información de las capas de ingreso implementando una función matemática previamente definida. Puede existir muchas capas ocultas, las cuales se conectan entre sí de diversas maneras.
- Capas de salida: Esta capa recibe las salidas de las capas ocultas y las lleva a un receptor externo

Estas características de las ANN las hacen apropiadas para diversas aplicaciones donde no se dispone de un modelo indetificable en un comienzo, como por ejemplo problemas de clasificación, el descubrimiento de patrones de fraude económico o la predicción de series temporales[23].

Capítulo 3

Modelos de Medición de Influencia en Twitter

La medición de influencia en Twitter ha sido motivo de debate desde la misma implementación del servicio. En un comienzo, Twitter ordenaba a los usuarios tomando la influencia como el número de seguidores que tenía cada perfil. Con el paso del tiempo, este método comenzó a ganar algunos retractores[11] y comenzaron a surgir distintas metodologías para intentar caracterizar la influencia en Twitter.

En el siguiente capítulo se mencionan diversos modelos, ordenados por fecha de publicación, que caracterizan la influencia en Twitter, para luego ver el que será usado posteriormente.

3.1. TwitterRank

TwitterRank [84] es un algoritmo publicado el 2010 que presenta uno de los primeros enfoques para medir la influencia en Twitter. Basado en PageRank[56], toma como primicia la posible existencia de homofilia dentro de la red social, lo que significaría que la conexión entre dos usuarios está muy relacionada por temas de interés.

3.1.1. Definición de Influencia

En este trabajo, un usuario influyente es aquel que tiene cierto nivel de autoridad en la red social. Esta autoridad puede distribuirse en distintos tópicos en los que se maneja un usuario, ya que no necesariamente todos tienen la misma influencia en todos los tópicos que manejan.

3.1.2. Set de Datos

Los autores recolectaron a los 1000 twitteros más influyentes de Singapur según la página twitterholic.com (que ordena por número de seguidores). De estos mil usuarios cuatro no estaban disponibles, por lo que se tomaron 996 twitteros y se extrajeron sus seguidores y los usuarios que seguían. De la unión de estos conjuntos se obtuvieron los usuarios de Singapur para finalmente extraer 3200 tweets de cada usuarios. No se pudieron extraer más tweets dado que es el mayor número de publicaciones que muestra Twitter en la línea de tiempo de un perfil y la máxima cantidad que se puede obtener desde la REST API de Twitter, por lo que finalmente los autores recolectaron 1,021,039 tweets.

3.1.3. Tópicos

La manera de identificar tópicos en TwitterRank se basa en la utilización de LDA. Aunque la manera más natural de representar cada documento sea a través de un tweet, aquí los autores agregaron la información de cada usuario para crear un documento que corresponde escencialmente a cada perfil de Twitter. Esto se realizó con la idea de poder determinar los tópicos en los cuales está interesado cada usuario.

3.1.4. Métricas

Cada documento, es decir cada usuario, tiene asociada tres matrices, estas son:

- Una matriz DT que representan la cantidad de veces que un palabra w de un usuario s ha sido asignada a un tópico t
- Una segunda matriz WT que representa el número de veces donde una palabra única w ha sido asgindad a un tópico t.
- Finalmente, un vector Z que representa el tópico t que fue asignado a la palabra w

Con estos valores se construye la similitud de los tópicos que existe entre dos usuarios. Al igual que en PageRank, los usuarios son conectados por la existencia de una conexión entre ellos y cuando se navega en la red se crea un vector de *salto* de un perfil sobre otro aleatorio, para así no quedar en una subred indefinidamente.

Obteniendo las probabilidades de transisión de un usuario a otro se calcula el puntaje en un tópico en específico, para luego agregar todos estos valores y obtener el puntaje general de influencia de un usuario.

3.1.5. Conclusiones

Los autores concluyen que es posible que exista homofilia en Twitter, es decir las personas se siguen por el interés que existe entre ellas. Gracias a ello , se tiene que el modelo repre-

senta de mejor manera la influencia que solo tomar la cantidad de seguidores de un usuario seguidores.

3.2. Trabajo de Cha *et al.*

Esta investigación[11] que data del 2010, intenta tomar la mirada de [17] sobre la difusión y adopción de influencias, es decir que es más importante una recomendación de un cercano y se debe tomar en cuenta el ritmo de adopción de una innovación de la sociedad antes de gastar grandes sumas de dinero en campañas de marketing. Con esta mirada intenta responder ciertas preguntas del marketing de ese entonces, como ¿Qué tan efectivas son las campañas de marketing que intentan influir? ¿Puede la influencia de la persona de una área ser trasferida a otra?

3.2.1. Definición de influencia

Se ve la influencia como la capacidad de provocar una acción en otros usuarios. Esto los autores lo traducen en tres distintas métricas para ver que tanto se relacionan los perfiles con provocar alguna reacción en la red social.

3.2.2. Set de datos

Los autores recolectaron datos de Twitter a gran escala. Luego de una autorización de la empresa obtuvieron todas las cuentas activas que tenían un identificador desde 0 hasta 80 millones, lo que les llevo a recolectar más de 50 millones de usuarios. Para tomar a los usuarios activos, que serían los analizados, ignoraron a aquellos que tenian menos de 10 tweets durante toda su existencia y aquellas que contaban con un nombre de usuario no válido. Finalmente obtuvieron un poco más de 6 millones de perfiles activos y el estudio se centró en como el entero set de más de 50 millones de usuarios interactuaban con el set de usuarios activos.

3.2.3. Tópicos

Para identificar tópicos los autores tomaron palabras claves relacionadas con diversos temas, como Michael Jackson con las palabras claves *Michael Jackson* y *#mj*, estudiando de esta manera los usuarios que participan y los tweets que son generados por estas palabras claves.

3.2.4. Métricas

Este trabajo toma en cuenta ciertas características de Twitter para realizar comparaciones de influencia, las cuales son:

- Seguidores: La cantidad de usuarios que sigue un perfil.
- Retweets: El número de veces que un tweet es *reenviado* por otros.
- Menciones: El número de veces que un usuario es mencionado por otros.

3.2.5. Conclusiones

El análisis de estas tres medidas llevaron a concluir que el número de seguidores no se relaciona con las demás, concluyendo que seguidores no aseguran mayor influencia, ya que no se genera actividad en la red social. Sobre la influencia por tópicos los autores concluyen que un usuario es influyente para una gran cantidad de tópicos y que esta influencia puede ser pasada a otros temas de manera más efectiva que esperar a un nuevo lider en cierto tópico.

3.3. Modelo de Bakshy *et al.*

El trabajo de Bakshy *et al.*[4] fue publicado el 2011 y se basa en un análisis de la difusión de información generada por tweets que contienen una URL en específico. Esto se realiza para determinar si es más rentable captar a un usuario influyente para generar cascadas de información o si vale la pena utilizar a pequeños grupos de personas para la difusión de información boca-a-boca.

3.3.1. Definición de Influencia

La influencia, al igual que en otros trabajos, se ve como la habilidad de afectar a otros. En el caso de este trabajo los autores ven esta habilidad por medio de la propagación que efectúa cierta URL en la red, desde que nace en un primer tweet hasta que termina de propagarse.

3.3.2. Datos

En un período de dos meses se obtuvieron todos los tweets públicos y se filtraron los Tweets con bit.ly URLs y se mantuvieron las URL que tenían el tweet *semilla*, o inicial, dentro del período de los dos meses obteniendo un total de 74 millones de tweets.

3.3.3. Tópicos

Para poder diferenciar los tópicos de los cuales las URLs están asociados, los autores usaron Amazon Turk Machine[1] para poder asociar cada URL a una de las 10 categorías que se establecieron de antemano.

3.3.4. Métricas

Los autores tomaron en consideración la cantidad de nodos por la que pasó una URL, asignando un puntaje de acuedo a si el perfil C vio primero al perfil A o al perfil B. La asginación de puntajes se hizo de tres maneras: la primera en donde el primer usuario que realiza el mensaje da el puntaje, la segunda donde el puntaje se divide en los perfiles A y B, y la tercera donde el último usuario es quien recibe el puntaje.

3.3.5. Conclusiones

Este modelo utilizó URLs para medir el movimiento dado que al momento de la investigación no existían los retweets. Lamentablemente, la sola utilización de URLs no implica que usuario es más influyente o no, sino más bien que URL es más interesante que otra, por lo que se obtuvo mucha influencia para usuarios que usaban URLs de videos o contenido de farándula.

3.4. ProfileRank

Este modelo [70]pretende encontrar perfiles influyentes en Twitter usando la estructura del perfil, la similitud de los tópicos entre los usuarios y la propagación de información.

Toma como primicia que el contenido relevante es creado y propagado por usuarios relevantes, y que usuarios relevantes generan nuevo contenido relevante.

3.4.1. Definición de Influencia

ProfileRank toma dos conceptos claves: *influencia* y *relevancia*.

Influencia es vista desde el enfoque de difusión de contenidos. Un contenido que es relevante es aquel que tiene un largo alcance luego de ser introducido por un usuario, y por ello un usuario influyente es aquel que produce contenido que es relevante para un grupo significado

[1] Amazon Mechanical Turk es un servicio de Amazon que permite utilizar inteligencia humana para ciertas tareas, así uno puede pagar cierta cantidad de dinero por realizar una acción y pedir que miles de personas la realicen para poder obtener resultados humanos.

de una comunidad. Esto se traduce como el alcance que tiene un contenido en la red social. Además, ProfileRank pretende dar recomendaciones personalizadas a cada usuario (en base a lo que él considera relevante) más allá de generar un recomendador global.

3.4.2. Set de Datos y Tópicos

Los autores recopilaron distintos períodos de Twitter para guardarlos en diferentes set de datos. Estos datos están relacionados con los autos, la liga brasileña y las elecciones de EEUU. No se explica en el trabajo el cómo se separaron los tópicos (es decir si utilizaron palabras claves para recopilar la información y cuales son estas palabras claves). También se utilizó un set de 17 millones de usuarios de un trabajo anterior de los autores y a través de la web Meme Tracker, se obtuvo un set aún mayor de como una frase se propagaba en el tiempo en Twitter.

3.4.3. Métricas

Como se tiene el enfoque de difusión de información se toma en consideración que un grafo de difusión de información es un grafo bipartito $G(U, C, F, E)$ donde U es el set de usuarios, C el set de contenido y E y F son los ejes asociados a los usuarios y al contenido. Es decir, para cada usuario $u \in U$ y cada pieza de información $c \in C$ hay un par $(u, c) \in E$ si el usuario u ha creado o propagado el contenido c y un par $(c, u) \in F$ si c fue creado por u.

A partir de esta definición se establece que el grafo G se puede representar como una matriz de usuario-contenido M y una matriz contenido-usuario L. En este caso la matriz $M = (m_{\mathrm{i},j})$ es una matriz $|U| \times |C|$ donde $m_{\mathrm{i},j} = 1/q_{\mathrm{i}}$ donde q_{i} es la cantidad de contenido que u_{i} ha generado o propagado. Así, $L = (l_{\mathrm{i},j})$ es una matriz $|C| \times |U|$ donde $l_{\mathrm{i},j} = 1$ si el usuario u_j creo el contenido c_{i} y 0 si es que no.

Basándose en tales matrices se define el contenido relevante y la influencia de un usuario como:

$$r = \mathrm{i}M \qquad\qquad \mathrm{i} = rL$$

Para estas dos ecuaciones debería conocerse el vector r o el vector i para que tengan solución, pero se pueden calcular de manera recursiva.

$$r^k = r^{k-1}LM \qquad\qquad \mathrm{i}^k = \mathrm{i}^{k-1}LM$$

Donde $k \geq 0$, r^0 y i^0 son vectores uniformes.

Para evitar problemas de subgrafos que están fuertemente conectados y que a un usuario le cueste salir del subgrafo se crea un factor d que denota una pequeña probabilidad de que el usuario pase de un usuario a otro aleatoriamente. Con la inclusión de este factor d se tiene:

$$r^k = \mathrm{d}r^{k-1}LM + (1 - \mathrm{d})u \tag{3.4.1}$$

$$\mathrm{i}^k = \mathrm{d}\mathrm{i}^{k-1}LM + (1-\mathrm{d})u \tag{3.4.2}$$

Donde u es un vector uniforme. Finalmente se puede reformular tal ecuaciones para obtener una manera no recursa de medir la influencia

$$r = (1-\mathrm{d})u(I - \mathrm{d}LM)^{-1} \tag{3.4.3}$$

$$\mathrm{i} = (1-\mathrm{d})u(I - \mathrm{d}LM)^{-1} \tag{3.4.4}$$

Donde I es la matriz identidad.

Cabe destacar que una versión *open-source* de este modelamiento se encuentra disponible en la red[2].

3.4.4. Conclusiones

El modelo de ProfileRank, según los autores, es una mejora sobre los modelos bases que se usaron. Se generan buenas recomendaciones para los usuarios y se encuentran perfiles influyentes en base a lo relevante del contenido que generan.

3.5. Trend Sensitive - LDA

Dado el auge de las redes sociales y la exorbitante cantidad de información presente en Twitter, surge el problema de poder saber qué es *interesante* para los usuarios y así tener una aproximación de cuáles son las necesidades de los usuarios en diverso ámbitos. Para ello los autores proponen TS - LDA[86], el cual es un modelo que pretende catalogar Tweets de acuerdo al nivel de interés que generan entre los usuarios de la plataforma, realizando también un análisis de tópicos.

3.5.1. Definición de Influencia

En este trabajo no se ve el nivel de influencia de un usuario, por su parte se ve el grado de interés que genera un tweet. El graod de interés se ve como el contenido que es de potencial interés para una larga audiencia. Así, por ejemplo, un tweet de Justin Bieber es solo importante para su nicho de seguidores y por ello no sería interesante para una gran parte de la audiencia de Twitter.

3.5.2. Set de Datos

En un período de 4 semanas se extrajeron los tweets, eliminando respuestas, retweets, URLs y tweets no ingleses, quedando una cantidad de 79.6 millones de tweets para entrenar

[2] `https://code.google.com/p/profilerank/`

el modelo. Al set resultante se le eliminaron *hashtags*, *stopwords* y se realizó *stemming* a las palabras restantes.

3.5.3. Tópicos

Al ser una extensión de LDA, los tópicos son generados por el modelo LDA. Este fue aplicado a 1.55 millones de tweets escogidos de manera aleatoria de 31 días sacando 50 mil tweets para cada día.

3.5.4. Métricas

Para poder medir el nivel de interés de un tweet los autores definen ciertos parámetros a considerar en base a observaciones realizadas. En LDA existen dos tipos de probabilidades de distribución: la probabilidad de la palabra w de ocurrir en el tópico t, es decir $p(w|t)$, y la probabilidad del tópico t de ocurrir en el documento d, es decir $p(t|\mathrm{d})$. Ambas probabilidades son consecuencias de LDA y en base a observaciones de ambas probabilidades se definieron los siguientes parámetros.

- Integridad de un tópico:
  Esto asume que no todos los tópicos son útiles para poder analizar la data, y es medido a partir de las palabras significantes de cada tópico. Muchos estudios que toman las palabras con mayor probabilidad de un tópico para poder representarlo. Esto lleva a que existan tópicos con palabras que no aportan al estudio, como en este caso palabras no inglesas o inusuales, por lo que los investigadores armaron un diccionario propio con palabras significativas, tomando como base el diccionario inglés y los nombres propios con una frecuencia mayor a 5.000 en un corpus de 13.5 millones de documentos. La integridad queda como:
  $$I(t) = \sum_{w \in W} p(w|t) L(w) \tag{3.5.1}$$
  Donde $p(w|t)$ es la probabilidad de una palabra w dado el tópico t y $L(w)$ es una variable que vale 1 si el diccionario generado contiene la palabra w y 0 si no la contiene.

- Entropía espacial:
  Este puntaje explota la observación de que los tópicos más significativos están asociado a un pequeño número de documentos, es decir los tópicos muy generales (con palabras como *hola* o *dia*) se intentan descartar. La entropía espacial se define como:
  $$S(t) = -\sum_{\mathrm{d} \in D} p(\mathrm{d}|t) log p(\mathrm{d}|w) \tag{3.5.2}$$
  Donde un documento d es un solo tweet (a diferencia de ProfileRank donde es todo un usuario) y $p(\mathrm{d}|t)$ denota la probabilidad del documento d dado un tópico t. Para obtener $p(\mathrm{d}|t)$ se utiliza inferencia bayesiana sobre $p(t|\mathrm{d})$.

- Entropía temporal:
  Representa la distinción de tópicos basado en la distribución de tópicos para un periodo específico. Al contrario de la *Entropía Espacial*, detecta los cambios de tópicos en Twitter y utiliza un set de tweets que tienen la misma fecha, ya que un solo tweet no tiene tanta información para encontrar los tópicos distintivos. Este score explota la observación de que los tópicos más significativos están relacionados a un específico periodo de tiempo, si se tiene que un tópico se mantiene en el tiempo es más posible que sea uno más general. La entropía temporal queda como:

  $$T(t) = -\sum_{s \in S} p(s|t) log p(s|t) \tag{3.5.3}$$

  Donde $p(s|t)$ denota la probabilidad aprendida de un tiempo s dado un tópico t y es medida usando $p(\mathrm{d}|t)$, esto es $p(s|t) = \sum_{\mathrm{d} \in D} p(\mathrm{d}|t)$, donde s equivale a un periodo de tiempo en específico, en este caso 1 día.
- Tópicos interesantes:
  Se establece un puntaje final luego de normalizar los puntajes descritos anteriormente. La importancia del tópico es representada por el peso de los tópicos latentes para medir el grado de interés de un solo tweet. Este peso es:

  $$W(t) = \tilde{I}(t) - \tilde{S}(t) - \tilde{T}(t) \tag{3.5.4}$$

  A la intregidad se le restan las entropía, ya que como se mencionó antes, las entropías indican la prescencia del un tópico repetidamente en el corpus o en el tiempo lo que para los autores significan temas menos interesantes.
- Tweets interesantes:
  Luego de tener ya el score de interés en un tópico se pasa a ver el score de interés de un solo tweet, esto a partir de suma de las probabilidades de un tópico dado un tweet multiplicado el score de tal tópico. Es decir:

  $$Score(\mathrm{d}) = \sum_{t \in T} W(t) p(t|\mathrm{d}) \tag{3.5.5}$$

  Un tweet va a generar más score si cubre tópicos latentes que tiene un alto puntaje. Así, se puede ordenar los tweet de acuerdo a su score de interés.

3.5.5. Conclusiones

Para poder comprobar los resultados de los tweets interesantes, los autores recurrieron al servicio de Amazon Turk Machine, donde pidieron a un grupo de personas que clasificara una gran cantidad de tweets para ver si son interesantes o no. Luego de tal clasificación se eligieron siete personas para poder comparar sus respuestas, así asignando un umbral para el interés se llega a la conclusión de que un tweet es interesante si al menos 3 personas lo marcaron como tal.

Con un set de contraparte, los autores pudieron comprobar su modelo frente a otras metodologías para cubrir tweets interesante, logrando que el modelo de TS-LDA superara a todos los demás.

3.6. Observaciones y elección

La primera observación que se puede realizar es las distintas maneras que existen para poder medir la influencia en Twitter. No existe una definición fija y establecida en como se entiende la influencia y solo se pretende llegar a resultados que no produzcan gran extrañeza y sean relativamente coherentes con la realidad, es decir, no llegar a la conclusión que alguien que solo ha publicado un tweet que nunca fue retweeteado pueda ser considerado muy influyente.

Segundo, la cantidad de seguidores y de cuentas que se siguen no son indicadores válidos de influencia ya que no muestran que exista un real movimiento en la red. Además, un indicador como este se puede prestar a mal utilización por parte de servicios que se dediquen a crear cuentas *spam* que solo se sigan personas o escriban tweets de manera automática. Por lo mismo, la cantidad de tweets escritos no es un indicador vláido.

Tercero, es esperable que el contenido relevante sea creado por personas importantes, y que este contenido relevante también sea interesante. Es por ello que los modelos más interesantes son ProfileRank y TS - LDA, donde el último puede ser un complemento a la idea central del primero ya que eliminaría el problema de obtener recursivamente los valores de los vectores de i y r. Por ello TS - LDA daría una primera aproximación a i para así calcular r.

Finalmente, se toma en consideración el último modelo presentado, el cual se ajusta a la hipótesis de que se puede generar un score de influencia en base al contenido interesante que generan las personas en Twitter. La aplicación de este modelo para medir a un usuario es vista en el capítulo siguiente.

Capítulo 4

Modelamiento de Influencia en Twitter

El siguiente capítulo tratará del modelamiento de la influencia en la red social de Twitter. Como se ha mencionado con anterioridad, la influencia es de difícil definición y no existe una respuesta objetiva a como se ve. Es por ello que se analizaron distintos datos presentes en Twitter siguiendo la presencia de que la influencia se puede considerar como *la manera de crear contenido interesante que repercute en la red* por parte de los usuarios.

En primera instacia se hablará de las métricas a usar. Posteriormente, se abordará la extracción de datos de Twitter. Finalmente, se hablará sobre como los datos obtenidos son tratados para poder llegar a las métricas propuestas.

4.1. Métricas

Dados los datos que pueden ser obtenidos en Twitter, más lo observado en los distintos modelos, se tiene que es necesario tomar en cuenta ciertas características de Twitter.

En primera instancia se toma en consideración el valor de lo interesante de un tópico de acuerdo a lo visto por TS - LDFA. Esto se realiza para lidar con la observación de [66] donde la mayoría de los temas en los que se habla en Twitter no tienen mayor trascendencia. Se tiene por lo tanto que un tópico interesante obtiene un valor identico al de la ecuación 3.5.4 en TS - LDA:

$$W(t) = \tilde{I}(t) - \tilde{S}(t) - \tilde{T}(t) \tag{4.1.1}$$

Por su parte, dado que se aplica el modelo de LDA, es posible obtener la probabilidad de un tópico dado un usuario, es decir $p(t|u)$ agregando la información de cada documento que está asociado al usuario, esta es $p(t|\mathrm{d})$. Así, se puede obtener un puntaje de lo interesante que es cada usuario:

$$UW(u) = \sum_{t \in T} p(t|u) W(t) \tag{4.1.2}$$

Esto podría no ser suficiente, y como se vió en los distintos modelos del capítulo anterior, la repercusión que genera un usuario en la red es una buena medida, por ello se obtiene además el número de retweets:

$$R(\mathrm{d}) = N\acute{u}mero \text{ de } veces\ que \text{ d } ha\ sido\ retweeteado \tag{4.1.3}$$

Pero como es posible que un usuario escriba muchas veces, se pondera el número de retweets por la cantidad de tweets que ha generado el usuario.

$$RT\_CT(u) = \frac{\sum_{\mathrm{d}\in U} R(\mathrm{d})}{\sum_{\mathrm{d}\in U} 1} \tag{4.1.4}$$

donde $\mathrm{d} \in U$ represanta los tweets que pertenecen a un usuario u.

Como también es importante la participación del usuario en la red social, este valor se multiplica por el logaritmo de la cantidad de tweets que ha generado más 1. Se multiplica por el logaritmo para darle más importancia a la cantidad de retweets por número de tweets que a la cantidad de tweets escritos, y el 1 es sumado para no provocar que el puntaje se vuelva 0 cuando existe un solo tweet.

$$S\_RT(u) = RT\_CT(u) * ln(1 + \sum_{\mathrm{d}\in U} 1) \tag{4.1.5}$$

Con este último puntaje y normalizando los valores de UW, se puede obtener un *score* de influencia para cada usuario de acuerdo a lo interesante de sus tweets.

$$IU(u) = S\_RT(u) * \widetilde{UW}(u) \tag{4.1.6}$$

Para obtener un puntaje diferenciado por temas, se obtiene un puntaje UW' el cual sea una suma de los tópicos que están relacionados a un tema y no de la suma de todos los tópicos, logrando un puntaje de interés sobre un nicho más específico.

4.2. Extracción de datos

La cantidad de datos en Twitter es exorbitante, con más de 250 millones de usuarios activos al segundo trimestre del 2014 [81], la cantidad de tweets crece y crece. Es por ello que, tal como en [70, 86, 84, 4] es recomendable tener un filtro de la cantidad de usuarios para poder analizar. Tomando en cuenta que se requerían obtener cuentas chilenas que fueran activas se siguieron una serie de pasos que se detallan a continuación.

4.2.1. Filtros de Extracción

Los diversos filtros utilizados en este trabajo se realizaron para obtener un set de datos lo suficientemente significativo para realizar los análisis pertinentes. Siguiendo el caso de [84],

donde se utilizaron los 1000 primeros usuarios con más seguidores de Singapur y luego se obtuvieron los perfiles que seguían, se tomaron las 253 cuentas chilenas con más de cien mil seguidores al 9 de septiembre del 2013 según Radio Cooperativa [14].

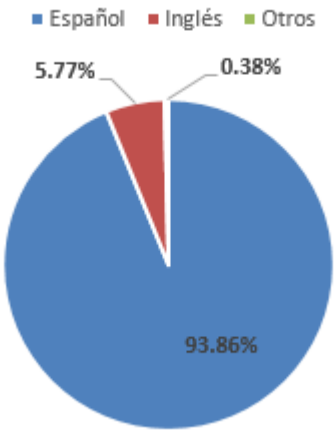
Luego, con ayuda de la librería Twitter4j[85][1], que hace uso de las API de Twitter [79, 80], se construyó un *crawler* en JAVA [54] con el cual es posible recopilar los seguidores de distintos usuarios para ser guardados. Así, comenzando con la lista de los usuarios con más de 100 mil seguidores se comenzaron a recopilar los distintos seguidores.

Para poder obtener un set representativo de datos chilenos y de cuentas activas, se realizaron diversos filtros que se describen a continuación.

- Primer filtro: Seguidores de las cuentas chilenas con más de cien mil seguidores
  Dado que el español es la segunda lengua con más usuarios en Twitter [51] obtener los usuarios activos de habla hispana llevaría a tener una altísima cantidad de información, es por ello que el primer filtro fue utilizar cuentas chilenas para la recopilación de seguidores, donde se usó las 253 cuentas con más de 100 mil seguidores al 9 de septiembre del 2013[14].
- Segundo filtro: Seguidores con cuentas en español
  Claramente los usuarios con más de 100 mil seguidores, sobre todo personajes de popularidad internacional como la cuenta de Bio Bio noticias, German Garmendia o el ex presidente Piñera, son seguidos por usuarios de distintos idiomas. En la figura 4.1 se puede apreciar los idiomas de los seguidores de BioBio, siendo casi todos en Español, por ello el segundo filtro a aplicar es que las cuentas a usar sean en español. Esto se logra con la misma información obtenida por el perfil de Twitter, donde se puede obtener el idioma de la cuenta y con ello ver si es equivalente al idioma deseado. Lamentablemente, existen usuarios que sus cuentas perfilan con otro idioma aun cuando sean chilenos, por lo cual estos usuarios son filtrados.
- Tercer filtro: Más de 90 seguidores
  Como se observa en la figura 4.2 la frecuencia de *followers* de los perfiles que siguen a BioBio disminuye drásticamente entre mayor sea el número de *followers*. Aunque el número de *followers* no indica un nivel de influencia en la red, si muestra cierto nivel de actividad dentro de ella dado que solo al participar de Twitter uno va adquiriéndolos sin solicitarlos. Es por ello que se tomó un umbral arbitrario para poder descartar a los usuarios con poca actividad, el cual fue fijado en que un perfil debe tener más de 90 seguidores.
- Cuarto filtro: Más de 20 tweets
  Tomando en cuenta la figura 4.2, se observa que el número de tweets que un perfil publica aumenta a medida que tiene más seguidores, es decir entre más activo es en la red, y dado que la cantidad de usuarios que solo utilizan el sistema para informarse[24] es bastante alta, y teniendo en cuenta que el fin del trabajo es encontrarse potenciales perfiles influyentes, se toman en consideración solo las cuentas que han realizado más de 20 publicaciones desde que fueron creadas, esto con el fin de poder tener usuarios activos.
- Quinto filtro: Si no está protegido

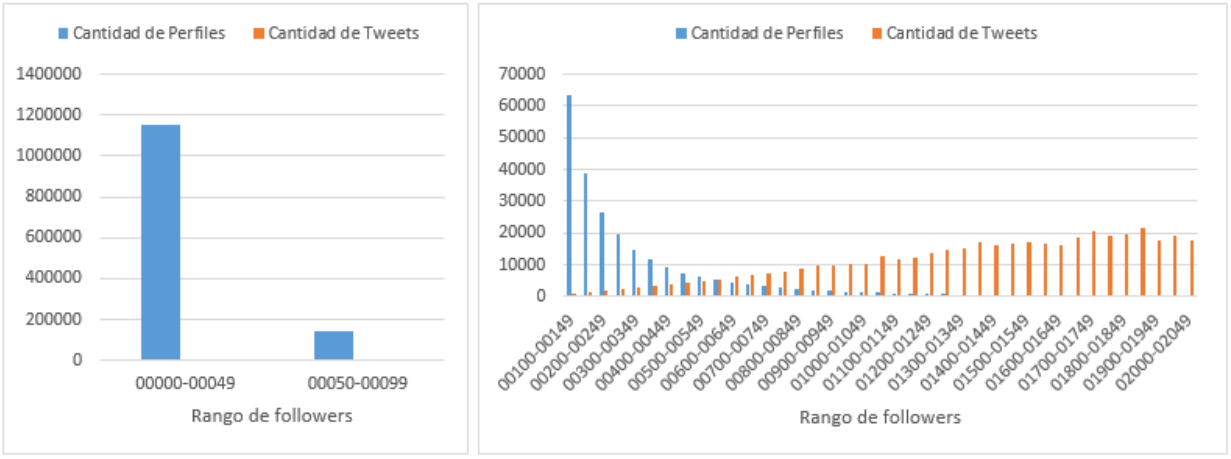
[1] http://twitter4j.org/en/index.html

Figura 4.1: Idioma de las cuentas de los seguidores de BioBio



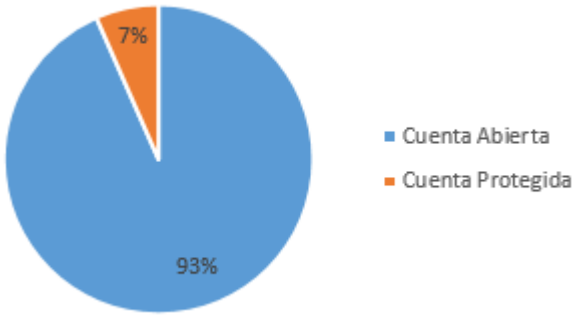
Fuente: Elaboración propia.

Figura 4.2: Cantidad de *followers* que tienen los seguidores de BioBio



Fuente: Elaboración propia, la figura de la derecha muestra los dos primeros rangos, mientras que la figura izquierda muestra hasta el rango de 2000 a 2049 seguidores. Esta diferencia se realizó dada la gran cantidad de perfiles que se encuentran en los primeros rangos.

Figura 4.3: Tipo de seguridad de cuenta de los seguidores de BioBio



Fuente: Elaboración propia.

Una cuenta protegida es aquella cuyos tweets no son públicos por lo que sus tweets no pueden ser analizados y no son de interés público. Para el caso de [84], existian alrededor del 10 % de perfiles con cuenta protegida. En el caso de los seguidores de BioBio, se puede observar en la figura 4.3 que exite un 7 % de perfiles que no dejan sus tweets al público. Esta cantidad es pequeña y estos perfiles no son de interés de estudio, por lo que son filtrados.

- Sexto filtro: El último mensaje tiene a lo más 15 días de antigüedad
  Como se pretenden utilizar cuentas activas, junto al cuarto filtro, es necesario tomar el cuenta que tan activo es un usuario, y para ello se utilizó al momento de la recopilación[2] que el último mensaje publicado en el perfil haya sido dentro de los últimos 15 días.
- Séptimo filtro: Su descripción no menciona ser de Chile
  Los perfiles de Twitter cuentan en su perfil una sección de localización en la cual se suele mencionar el lugar de procedencia. Dado que muchos usuarios no colocan su localización, si es el caso no son filtrados, pero aquellos que si colocan la localización se espera que mencionen a Chile o alguna ciudad del país, si no es el caso son filtrados. Hay perfiles que tienen en su localización lugares como "el mundo", "tu mirada", "en mi casa", y demáses, los cuales son filtrados de todos modos.

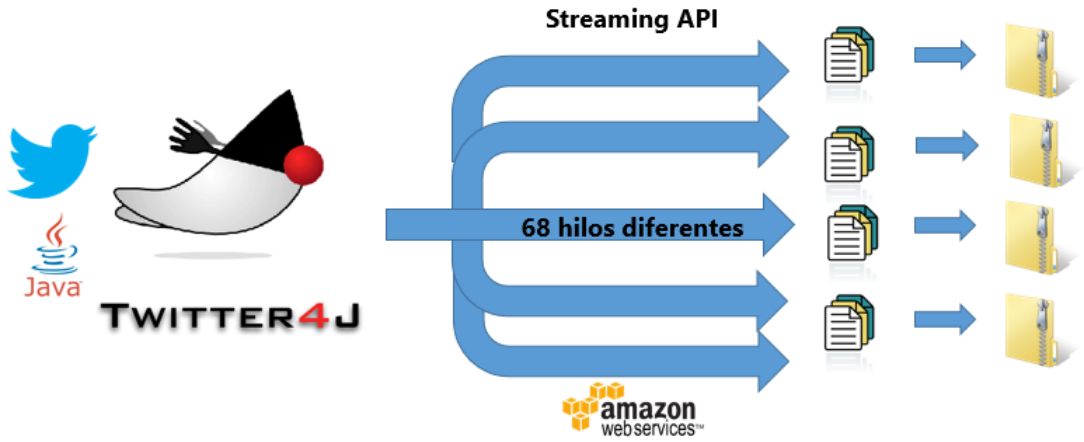
Finalizando los pasos anteriores se obtuvieron cerca de ∼335.000 usuarios a principios de septiembre del 2014 con los cuales trabajar.

Luego, con tal cantidad de usuarios es necesarios obtener información de los tweets que van generando los perfiles. Como se muestra en 4.4 con la Streaming API [79] se recopilaron simultáneamente los tweets que iban generando los set de usuarios. Cada tweet es recopilado en formato JSON, logrando una gran cantidad de información. Por ello, cada vez que se generaban 5.000 tweets en un set de usuarios esta información era comprimida en formato zip[68] con la ayuda de la libreria zip de JAVA, logrando un ratio de compresión cercano al 95 %.

Los tweets fueron recopilados desde el 17 de diciembre del 2014 hasta mediados de febrero

[2]Dado que la recopilación de usuarios fue durante varios días, esta fecha es variable.

Figura 4.4: Obtención de datos a través de la Streaming API



Fuente: Elaboración propia.

del 2015.

4.3. Tópicos

Dado que se usa TS-LDA[86], el módelo de tópicos a usar es Latent Dirichlet Allocation. Existen diversas librerias para poder aplicar este modelo, como por ejemplo Mallet[45][3] o JGibbLDA[59][4]. para JAVA.

La importancia de diferenciar por tópicos en Twitter (o cualquier red social) radica en que los usuarios no son necesariamente expertos en variados temas y, por lo tanto, no tendrían por qué ser influyentes en todos ellos. Un estudio de Pearanalitycs [66] llegó a la conclusión de que el 40 % de los tweets son *cháchara* sin sentido y un 38 % conversaciones casuales, dejando un pequeño porcentaje para otros tipos de tweets. Dado estos porcentajes, si no se dividiera la información por tópicos se tendría una gran cantidad de datos que serían ruido para el análisis. Es por ello que TS - LDA presenta un atractivo enfoque para resolver este problema quitando peso a tópicos que se mantienen en el tiempo o que están presentes en un largo número de documentos.

[3]MALLET es un paquete basado en Java para diversos usos de text mining, como por ejemplo estadísticas natural del lenguage, clasificación de documentos, clusterización, modelamiento de tópiocs, extracción de información y otras aplicaciones de machine learning a texto. URL: http://mallet.cs.umass.edu/

[4]JGibbLDA es una implementación de LDA en Java usando muestreo Gibbs para la estimación de parámetros e inferencia URL: http://jgibblda.sourceforge.net/

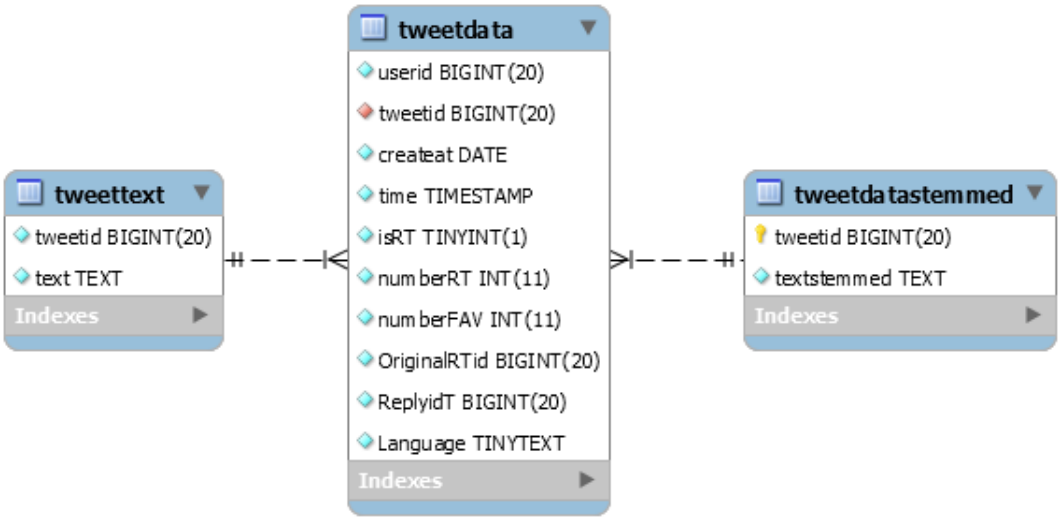
4.4. Aplicación

Los puntajes a obtener han de utilizar TS-LDA y el número de RT por cantidad de tweets. Toda la información anterior se encuentra en las líneas de texto en JSON obtenidas del *streaming* realizado a los usuarios. Esta información tiene que ser preparada y luego aplicada con diferentes algoritmos para poder obtener el resultado deseado.

4.4.1. Preparación de la data

Teniendo ya los JSON de los usuarios se procedió a rescatar la siguiente información de cada string en una base de datos relacional en PostgreSQL[62], cuyo contenido se detalla en la figura 4.5:

Figura 4.5: Base de datos con datos para la medición de influencia



Fuente: Elaboración propia.

- tweettext:
  La finalidad de esta tabla es tener el texto de cada tweet de manera separada a la información general, para ello se tienen dos variables:
  - tweetid:
    El identificador de cada tweet
  - text:
    El contenido textual de cada tweet
- tweetdata:
  La finalidad de esta tabla es tener la información extra de cada tweet de manera separada del texto, para ello se tienen las siguientes variables variables:
  - userid:
    El identificador del creador del tweet

  - tweetid:
    El identeificado de cada tweet
  - createat:
    La fecha, con día, mes y año, de la creación del tweet
  - time:
    La estampa del tiempo del tweet, es igual a la anterior más la hora y los minutos
  - isRT:
    Una variable que es 1 o 0 dependiendo si es un retweet o no
  - numberRT:
    La cantidad de veces que el tweet ha sido retweeteado
  - numberFAV:
    La cantidad de veces que el tweet ha sido marcado como favorito por otros usuarios
  - OriginalRTid:
    La id del tweet original si fue retweeteado, si es original o el tweet no fue retweeato directamente desde la función de Twitter este valor es -1
  - ReplyidT:
    El identificador del tweet al que el actual tweet es respuesta, el valor es -1 si es que no es una respuesta
  - Language:
    El lenguaje del tweet[5]
- tweetdatastemmed:
  La finalidad de esta tabla es tener el texto de cada tweet procesado para su comparación con el texto normal, para ello se tienen dos variables:
  - tweetid:
    Al igual que todas las tablas anteriores, el identificador de un tweet
  - textstemmed:
    La información textual de un tweet ya procesada para su uso posterior

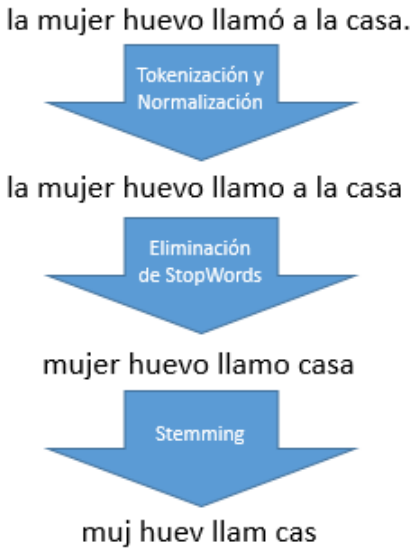
Como se observa, existe una tabla llamda *tweetdatastemmed* que contiene la información textual de cada tweet procesada por distintos pasos (ver esquema y ejemplo en la figura 4.6). La tokenización y normalización fueron realizadas en JAVA para poder tener cada palabra sin ninguna tilde; así quedan solo palabras y espacios vacíos en cada línea de texto.

Posteriormente se procede a eliminar las Stopwords. Para ello se utilizó la lista de palabras del apéndice B que es una mezcla de dos set de datos[6]. Como se mencionó en el capítulo 2, la selección de *stopwords* no es algo objetivo y en español usualmente son cojunciones, determinantes y conjunciones. Para eliminar estas stopwords de cada línea de texto, se ve la cadena de texto completa y se revisa cada stopword para que no esté presente en la cadena de texto, para finalmente devolver la cadena de texto sin alguna stopword.

[5]Cada usuario no elige el idioma en el que fue escrito su tweet, sino que Twitter intenta determinar de la mejor manera el idioma en el cual el tweet fue escrito.

[6]El primer set de *stopwords* es del trabajo del lingüista Sadowsky que generó una lista de palabras usadas en los medios chilenos. Este set de palabras se encuentra lematizado, por lo cual se utilizó un segundo conjunto de *stopwords* para complementar con las diferencias de género de algunos lemas. Este segundo set fue obtenido de http://www.ranks.nl/stopwords/spanish, una página de un servicio neerlandés de Herramientas Análiticas de Palabras Claves de páginas web con más de 15 años de experiencia.

Figura 4.6: Pasos para el stemming de los datos



Fuente: Elaboración propia.

Finalmente, se aplica el algoritmo Snowball de Porter para realizar el stemming de datos. Porter facilita las clases necesarias en diversos formatos para poder realizar este trabajo, así se toma cada token de la cadena de texto en JAVA y se obtiene la raíz de esta, guardándose en la base de datos.

4.4.2. Aplicación TS-LDA

Para poder obtener el puntaje de la ecuación 4.1.1 se necesita seguir los pasos del modelo TS - LDA tal como se especifican en el sub-capítulo 3.5. Para ello, se debe aplicar LDA y, luego, con las probabilidades calculadas, obtener los puntajes de la Integridad de un tópico y sus entropías espaciales y temporales. A continuación, se explica como esto fue obtenido.

Aplicación Latent Dirichlet Allocation

Como se mencionó en 4.3, se ha de usar LDA para obtener los tópicos, para ello se utilizó JGibbLDA. Esta librería necesita que los archivos de entrada estén en el siguiente formato:

$[M]$
$[\text{d}ocumento_1]$
$[\text{d}ocumento_2]$
...
$[\text{d}ocumento_M]$

En donde la primera línea es el número total de documentos $[M]$, y cada línea que sigue

es un documento [d$ocumento_{\mathrm{i}}$]. El i$-$$ésimo$ documento de un set de datos que tiene una lista de N palabras. Por lo que cada documeto tiene una estructura de:

$$[\mathrm{d}ocumento_{\mathrm{i}}] \quad = \quad [palabra_{\mathrm{i}}j] \quad [palabra_{\mathrm{i}}2] \quad ... \quad [palabra_{\mathrm{i}}N]$$

En donde cada palabra $[palabra_{\mathrm{i}}j] \quad (\mathrm{i} = 1, ..., M) \quad (j = 1, ..., N)$ son cadenas de texto separadas por un espacio en blanco.

Así, para poder llegar a tal archivo se extrajeron los datos en la base PostgreSQL, se contabilizaron la cantidad de registros y se creó un archivo donde la primera fila es el número de documentos y cada línea es un resultado de la consulta SQL de la base de datos explicada anteriormente.

Posteriormente, para poder aplicar el modelo es necesario ejecutarlo con ciertos parámetros. LDA pide los hiperparámetros α y β como también el número de tópicos que se quiere obtener. El número de tópicos no es un valor predefinido o fijo y usualmente se prueba el modelo con distintos valores hasta poder obtener un parámetro que parezca aceptable. Algunos trabajos usan el modelo HDP [75] para poder determinar la cantidad de grupos a usar en LDA.

Luego de ejecutar el modelo este entrega diversos archivos como resultados, los cuales son:

$< nombre - modelo > .others$
$< nombre - modelo > .phi$
$< nombre - modelo > .theta$
$< nombre - modelo > .tassign$
$< nombre - modelo > .twords$

Donde $< nombre - modelo >$ es el nombre del modelo correspondiente al número de iteración al que es guardado el archivo. Los demás archivos contienen la siguiente información:

Así, para poder determinar los tópicos se usaron los tweets recolectados por streaming durante cuatro semanas. Quitando retweets, respuestas y tweets con menos de 3 tokens quedó una cantidad de más de 6,000,000 de tweets. Al aplicar LDA se obtuvieron tópicos que no tenían que ver con Chile, por ello se decidió eliminar completamente a los usuarios que no especificaran que fueran de Chile y aquellos que tenían menos de 100 seguidores, o sea que su número de seguidos no aumente más de 10 puntos en más de 3 meses. Finalemnte quedo un total de 3,485,312 tweets, de los cuales se extrajeron una muestra para cada día. Como se observa en la figura 4.7 la cantidad de Tweets no es constante, por lo que se tomó la cantidad de un aproximadamente un 50 % para cada día, quedando un total de 1,742,656 de documentos para aplicar LDA.

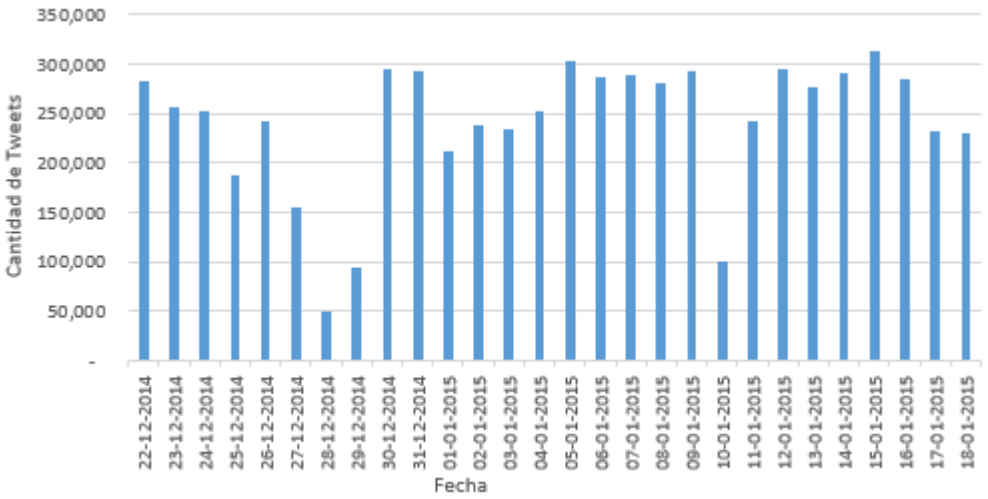
Se aplicaron 1,000 iteraciones de LDA tomando un α de $1/T$ y β de 0,1, donde T es el número de tópicos. El número de tópicos a trabajar fue de 200, esto basado en como fue elaborado TS-LDA[86] y en [89], mientras que los valores de α y β se basaron en [27]. Para poder ejecutar el modelo se usó Jgibblda en una instancia de Amazon Elastic Compute

Tabla 4.1: Archivos de salida de JGibbLDA

Nombre del archivo	Descripción
$< nombre - modelo > .others$	Este archivo contiene los diversos parámetros usados en el modelo LDA, como el valor de alpha, beta, el número de tópicos y el número de documentos.
$< nombre - modelo > .phi$	Este archivo contiene los valores de las distribuciones de cada palabra por tópico, es decir $p(palabra_w \| topico_t)$, siendo cada línea es un tópco y cada columna es una palabra.
$< nombre - modelo > .theta$	Este archivo contiene los valores de las distribuciones de cada tópico por documento, es decir $p(topico_t \| documento_m)$, siendo cada línea es un documento y cada columna es tópico probabilidad de aparecer en un tópico.
$< nombre - modelo > .tassign$	Este archivo contiene la asignación de un tópico para cada palabra, siendo cada línea documento consistente de $< palabra_{ij} >:< topicodelapalabraij >$.
$< nombre - modelo > .twords$	Este archivo contiene las palabra con mayor probabilidad de aparecer en un tópico, siendo la cantidad de palabras a mostrar un parámetro de entrada.

Fuente: Elaboración propia a partir de [59].

Figura 4.7: Cantidad de Tweets por fecha



Fuente: Elaboración propia. La gran diferencia que existen algunas fechas, como 28/12, es la paralización del *streaming* producto de llenarse el disco duro usado para la recolección de tweets.

Cloud[7] de 30gb de RAM dado lo grande de las matrices a usar (1,742,656 $\times$ 200).

Integridad de un Tópico

Para la integridad de un tópico, como se vió en el sub-capítulo 3.5, es necesario tener un diccionario de palabras que sean aceptadas para uso. Para ello se utilizó el corpus lingüístico LIFCACH (Lista de Frecuencias de Palabras del Castellano de Chile)[67]. Este corpus cuenta con la frencuencia de palabras de distintos documentos del habla chilena, donde se encuentran medios digitales, diarios, revistas, sitios gubernamentales, entre otros, inclusive a llegar a transcripciones orales de entrevistas y programas de televisión. Esta formado de 476.776 lemas, derivados de aproximadamente 4.5 millones de tipos presentes en 450 millones de palabras.

De este documento se extrajeron los lemas que cuentan con una frecuencia mayor a 1,000, dejando de lado las *stop words* del documento. Se asignaron como stopwords los adverbios, las conjunciones, los determinantes, las interjecciones, los pronombres y las preposiciones que tuvieran una frecuencia mayor a 50, quedando finalmente un diccionario de 19,399 lemas que pasaron por el algoritmo de Porter para extraer raíces. Cabe destacar que se agregaron algunos términos, como por ejemplo *hebd* de *Hebdo*, por los acontecimientos ocurridos a la revista satírica Charlie Hebdo el 7 de Enero del 2015.

Finalmente, con este diccionario, se fue viendo que cada token del tweet procesado estuviera presente para ver su integridad.

Entropía espacial

La entropía espacial es, en términos simples, una medida que muestra que tanto un tópico se encuentra presente en muchos documentos, este puntaje se puede apreciar en la ecuación 3.5.2. Para poder obtener este valor es necesario obtener $p(\mathrm{d}|t)$ a partir de los resultados de LDA. Para medir $p(\mathrm{d}|t)$ se utiliza $p(\mathrm{d}|t) = p(\mathrm{d})p(t|\mathrm{d})/p(t)$ en base a la inferencia bayesiana. El valor de $p(\mathrm{d})$ es simplemente $1/D$, con D el número de documentos; el valor de $p(t)$ se obtiene a través de $p(t|\mathrm{d})$, sumando cada valor de t en base a todos los documentos d para luego obtener la proporción de $p(t)$; por su parte $p(t|\mathrm{d})$ es el valor de θ que se obtiene a través de Jgibblda como se vió en la sección 4.4.2.

Entropía temporal

La entropía temporal, al igual que la espacial, es una medida de dispersión siendo en este caso que tanto un tópico está presente en distintos días. La ecuación 3.5.3 muestra el puntaje que ha de ser obtenido. En este caso, cuando se preparan los archivos para ser usados en Jgibblda, también se crea un archivo que contiene el día de creación de cada documento. Por

[7] Amazon EC2: http://aws.amazon.com/es/ec2/

ello, $p(s|t)$ es calculado a partir de $p(\mathrm{d}|t)$ sumando estas probabilidades para cada documento que se encuentre en el día s.

4.4.3. Número de Retweets y favoritos

En el caso del número de Retweets y favoritos con los que cuenta cada tweet, son proporcionados por el mismo set de datos. Cada documento está asociado a la id que le asigna Twitter, así fue posible obtener la información del número de Retweets y Favoritos.

4.5. Resultados preliminares

Los resultados preliminares corresponden a la aplicación de LDA, de TS-LDA, de la clasificación de tweets, y por consiguiente, de la clasficación de usuarios de acuerdo a su nivel de influencia.

4.5.1. Resultado de LDA

Como se ha mencionado varias veces, LDA ofrece una manera de asignar a un número de tópicos determinado la probabilidad de las palabras que aparecen en los documentos de estar presentes en los tópicos. En este trabajo se realizaron 1.000 iteraciones para poder obtener los 200 tópicos que representan al set de datos. La lista de tópicos se encuentra en la tabla A.2 del apéndice.

La figura 4.8 muestra el etiquetado manual que se realizó a los tópicos obtenidos con LDA, donde destaca como los temas banales y sin mayor importancia (cháchara) son la gran parte de los tópicos que se habla en Twitter, llegando a casi un 50 % del total.

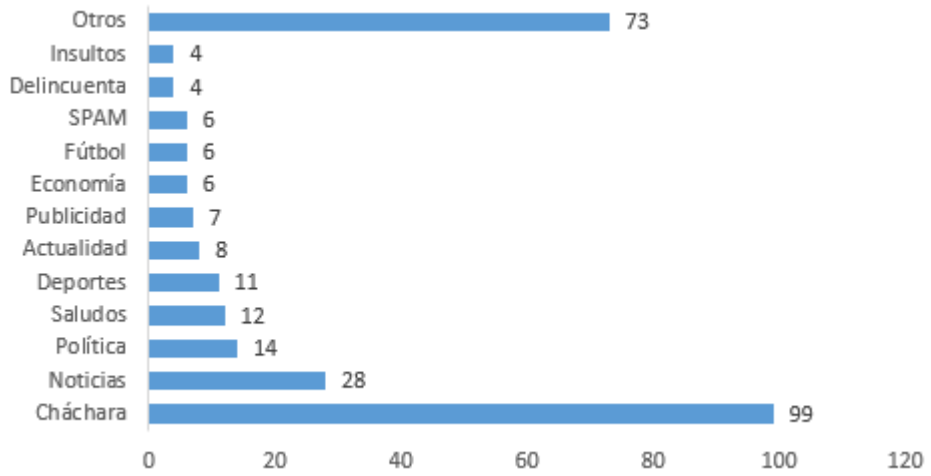
4.5.2. Resultado de TS-LDA

En cuanto a TS-LDA, se pudo obtener los tópicos *interesantes* dado las componentes de integridad, de entropía espacial y entropía temporal. Estas componentes muestran como un tema es más interesante que los demás.

Integridad

La integridad al estar representada como la suma de la probabilidad de que una palabra esté presente en un diccionario dado, se puede esperar que los temas que presentan una mejor escritura se encuentren mejor evaluados que aquellos cuya escritura no sea la más adecuada.

Figura 4.8: Temas de los tópicos de LDA



Fuente: Elaboración propia, el valor total supera los 200 tópicos dado que un tópico puede ser asignado a más de una etiqueta.

Tabla 4.2: Los 10 tópicos más y los 10 menos íntegros del 22/12/14 al 18/01/15

a) Primeros 10

Tópico	P. 1	P. 2	P. 3	P. 4	P. 5	P. 6	P. 7	P. 8	Integridad
78	vin	mar	diari	jardin	arriend	renac	gtgt	botanE	0.79
93	person	excelent	vin	vist	info	dept	renac	arriend	0.72
126	feliz	navid	cumplean	tod	famili	dese	les	des	0.69
155	incendi	bomber	forestal	sector	lug	alert	pastizal	roj	0.68
165	ano	nuev	feliz	sea	celebracion	exit	abraz	celebr	0.67
83	diput	reform	comision	senador	proyect	sistem	vot	binominal	0.67
109	sur	sector	rut	nort	pist	vehicul	km	accident	0.67
57	cas	pent	carl	declar	velasc	fiscali	andres	lavin	0.66
24	ano	nuev	fiest	celebr	empez	desped	proposit	comienz	0.65
47	plan	regional	alcald	intendent	realiz	entreg	comun	autor	0.64

b) Últimos 10

Tópico	P. 1	P. 2	P. 3	P. 4	P. 5	P. 6	P. 7	P. 8	Integridad
0	cos	hay	vec	much	tant	piens	cualqui	pas	0.35
114	the	gust	vide	of	trail	gam	and	hous	0.34
162	pregunt	dic	eso	dig	import	hic	respuest	dec	0.33
123	pued	nuev	vist	gtgt	verl	vinculocl	servidor	vps	0.33
46	wea	put	esa	mierd	weon	hue	wn	fom	0.32
16	quier	te	twitt	descubr	alta	tus	animal	dist	0.32
122	justin	one	sig	harry	re	niall	zayn	direction	0.29
139	wn	po	ctm	weon	xd	cag	oye	jajajaj	0.22
5	q	hay	dic	cre	dec	eso	xq	x	0.13
61	d	q	x	n	cn	t	ls	gob	0.06

Fuente: Elaboración propia.

Tabla 4.3: Los 10 tópicos con más y los 10 con menos entropía espacial del 22/12/14 al 18/01/15

a) Primeros 10

Tópico	P. 1	P. 2	P. 3	P. 4	P. 5	P. 6	P. 7	P. 8	E. Espacial
82	pas	mejor	vid	ide	pud	mia	import	dar	14.32
49	igual	sup	xd	razon	jaj	teni	vi	jajaj	14.32
183	just	podr	ver	ser	fras	necesari	termin	complet	14.31
184	person	esa	hay	dec	cre	palabr	sient	esas	14.31
63	volv	toc	pas	futur	dej	esper	congres	guitarr	14.31
135	com	hambr	uu	mayor	xd	jueg	dio	qued	14.31
180	tod	nuestr	famili	amig	apoy	quer	graci	compart	14.31
162	pregunt	dic	eso	dig	import	hic	respuest	dec	14.31
192	tod	lad	dias	junt	igual	fuerz	andan	vay	14.31
25	habl	dej	trat	corazon	eso	romp	entend	aprend	14.31

b) Últimos 10

Tópico	P. 1	P. 2	P. 3	P. 4	P. 5	P. 6	P. 7	P. 8	E. Espacial
65	atent	ataqu	terror	charli	franci	hebd	paris	muert	14.17
19	social	red	necesit	marc	gtgt	contactan	conoc	administr	14.17
109	sur	sector	rut	nort	pist	vehicul	km	accident	14.17
128	sig	te	vuelt	estas	ver	ofert	invit	twitt	14.11
155	incendi	bomber	forestal	sector	lug	alert	pastizal	roj	14.11
1	region	in	santiag	metropolitan	at	valparais	im	provident	14.09
123	pued	nuev	vist	gtgt	verl	vinculocl	servidor	vps	14.02
163	b	esq	fueg	llam	basur	artificial	r	hum	14.01
78	vin	mar	diari	jardin	arriend	renac	gtgt	botanE	13.59
93	person	excelent	vin	vist	info	dept	renac	arriend	13.56

Fuente: Elaboración propia.

La tabla 4.2 muestra como los primeros tópicos son fácilmente identificables, hablando de arriendos, relaciones exteriores, politicas, familia entre otros. Mientras que por otro lado los temas con peor integridad tienen palabras mal escritas y suelen asociarse a conversaciones banales que no son de interés público.

Entropía Espacial

La entropía espacial se refiere a la redundancia de temas a través de los distintos documentos del corpus intentando otorgar un gran puntaje a los temas que se repiten más dentro del corpus.

La tabla 4.3 muestra como los tópicos con mayor puntaje suelen ser de conversaciones comunes y corrientes lo cual es acorde a la realidad[66]. En el otro extremo existen temas que son mencioandos en menos documentos y con ello tienen una menor entropía espacial. Esto se realiza para poder resaltar temás que resultarían más interesantes de leer en el inicio de Twitter. Lamentablemente cuando pocos documentos se refieren a un tópico específico se

Tabla 4.4: Los 10 tópicos con más y los 10 con menos entropía temporal del 22/12/14 al 18/01/15

a) Primeros 10

Tópico	P. 1	P. 2	P. 3	P. 4	P. 5	P. 6	P. 7	P. 8	E. Espacial
104	estudi	univers	colegi	nacional	educacion	result	carrer	psu	3.29
166	gol	alexis	sanchez	arsenal	city	part	premi	golaz	3.29
106	u	part	azul	jug	pat	rubi	equip	enzo	3.28
36	via	amp	ft	by	descarg	prod	check	official	3.28
7	fot	sac	jav	sub	instagram	aceved	mostr	saqu	3.28
1	region	in	santiag	metropolitan	at	valparais	im	provident	3.28
127	fot	facebook	nuev	public	he	publiqu	album	set	3.28
163	b	esq	fueg	llam	basur	artificial	r	hum	3.28
100	encuentr	vuelv	loc	ciudadan	via	normal	volvi	ener	3.28
4	amig	mis	companer	favorit	tod	mejor	compart	secret	3.28

b) Últimos 10

Tópico	P. 1	P. 2	P. 3	P. 4	P. 5	P. 6	P. 7	P. 8	E. Espacial
83	diput	reform	comision	senador	proyect	sistem	vot	binominal	3.27
55	ena	error	von	baer	val	moreir	pid	ped	3.27
57	cas	pent	carl	declar	velasc	fiscali	andres	lavin	3.27
126	feliz	navid	cumplean	tod	famili	dese	les	des	3.27
65	atent	ataqu	terror	charli	franci	hebd	paris	muert	3.27
50	larrain	martin	justici	atropell	juici	fall	conden	mat	3.27
167	ministr	abort	molin	salud	renunci	heli	clinic	dich	3.26
123	pued	nuev	vist	gtgt	verl	vinculocl	servidor	vps	3.26
78	vin	mar	diari	jardin	arriend	renac	gtgt	botanE	3.25
93	person	excelent	vin	vist	info	dept	renac	arriend	3.18

Fuente: Elaboración propia.

genera una entropía espacial muy baja, lo que puede llevar a ciertas cuentas que realizan *spam*. Los últimos dos tópicos son referentes a arriendos en el litoral provocado por Tweets generados por unas pocas cuentas *spam*.

Entropía Temporal

La entropía temporal al referirse a la persistencia de los tópicos en tiempo dará un mayor puntaje a los tópicos que estén en varios documentos y que se repitan en el tiempo. Si pocos documentos presentan un tópico en un corto periodo de tiempo tendrán una menor entropía temporal, provocando que cuentas *spam* que generen mucho contenido durante varios días no tengo un puntaje muy bajo.

La tabla 4.4 muestra como los temas que son recurrentes tienen un puntaje mayor a aquellos que no lo son. A su vez, muestra como los dos últimos tópicos (que tratan de arriendos en el litoral) cuentan con una gran diferencia de entropía espacial siendo que son de cuentas *spam*, este suceso es tratado con mayor profundidad en la sección 6.1.2.

Tabla 4.5: Los 10 tópicos más y los 10 menos interesantes del 22/12/14 al 18/01/15

a) Primeros 10

Tópico	P. 1	P. 2	P. 3	P. 4	P. 5	P. 6	P. 7	P. 8	W
93	person	excelent	vin	vist	info	dept	renac	arriend	22.98
78	vin	mar	diari	jardin	arriend	renac	gtgt	botanE	14.72
126	feliz	navid	cumplean	tod	famili	dese	les	des	3.64
155	incendi	bomber	forestal	sector	lug	alert	pastizal	roj	3.57
167	ministr	abort	molin	salud	renunci	heli	clinic	dich	3.56
123	pued	nuev	vist	gtgt	verl	vinculocl	servidor	vps	3.44
163	b	esq	fueg	llam	basur	artificial	r	hum	3.42
65	atent	ataqu	terror	charli	franci	hebd	paris	muert	3.25
57	cas	pent	carl	declar	velasc	fiscali	andres	lavin	3.23
50	larrain	martin	justici	atropell	juici	fall	conden	mat	3.13

b) Últimos 10

Tópico	P. 1	P. 2	P. 3	P. 4	P. 5	P. 6	P. 7	P. 8	W
102	hrs	tod	pierd	tuit	quier	mont	music	ven	-2.07
16	quier	te	twitt	descubr	alta	tus	animal	dist	-2.16
89	gent	hay	odi	entiend	carg	cre	dic	esa	-2.16
0	cos	hay	vec	much	tant	piens	cualqui	pas	-2.41
46	wea	put	esa	mierd	weon	hue	wn	fom	-2.58
162	pregunt	dic	eso	dig	import	hic	respuest	dec	-2.7
122	justin	one	sig	harry	re	niall	zayn	direction	-2.81
139	wn	po	ctm	weon	xd	cag	oye	jajajaj	-3.82
5	q	hay	dic	cre	dec	eso	xq	x	-4.57
61	d	q	x	n	cn	t	ls	gob	-5.1

Fuente: Elaboración propia.

Puntaje de interés de un tópico

Normanlizando cada uno de los puntajes visto anteriormente y restando las entropías de la integridad se obtiene el puntaje de interés de un tópico (W).

La tabla 4.5 muestra, finalmente, el puntaje otorgado a cada uno de los tópicos. Se puede apreciar como en los primeros 10 temas existen sucesos específicos como el atentado a la revista satírica francesa Charlie Hedbo[52], la renuncia de la ministra de salud por sus dichos sobre el aborto[25], o el caso del cuestionado financiamiento político a través de las empresas Penta[12]. A su vez, los primeros dos tópicos son temas que presentan una entropía espacial y temporal muy baja provocando que su puntaje se aleje bastante del resto. Por lo anterior se truncó el valor de los dos tópicos *outliers* a un puntaje de 0.En los tópicos menos interesantes se presentan temas con palabras mayoritariamente mal escritas, lo que radica en un bajo puntaje.

Finalmente, se etiquetaron todos los tópicos con palabras claves (como política o deportes)

de acuerdo a las palabras que más los representan.

4.5.3. Retweets y Favoritos

La muestra para realizar TS-LDA no cuenta con el número de retweets, favoritos o si los tweets son respuestas, por lo que no es posible obtener directamente el número de retweets que se realizaron a los mensajes analizados. Es por ello que se toma el identificador de cada mensaje y se hace una búsqueda para ver si existe algún mensaje que retweeteó a éste. Así si un tweet ha sido favorito pero no retweeteado no puede ser contabilizado, ya que en la recolección de datos se toma el tweet cuando se genera, por lo que estos dos datos son nulos y solo se puede obtener el número de retweet y favoritos al momento en que aparece un mensaje que retweetea al original.

Tabla 4.6: Los 5 Tweets con más RT del 22/12/14 al 18/01/15

Cuenta	Tweet	N° de RT
MalaImagen	La viñeta de hoy. Comparta y/o comente si quiere. #CharlieHebdo http://t.co/aj7Vbn3b6c http://t.co/jtuiGf0Vnq	3798
C1audioBravo	Gran victoria, gran partido y muy bien jugado. Vamos por más!!! ??????????	3200
canal13	RT / Muchas gracias a todos por el cariño entregado durante estas 7 temporadas #HastaSiempre #Los80 :) http://t.co/yG Vv1LFK5r	1104
TaniaMelnick	Soy asesino de niños y de periodistas en Gaza... Y marcho por la unidad y en contra del terrorismo	1072
EnciclopediaCL	NO TE VEO DEL AÑO PASADO: frase muy aweoná que te dicen el 1 de Enero. Al weon que te diga esa wea hay que agarrarlo a balazos por WEON	1033
JorgeAlis	#CasoPentaTVN marcando 5 puntos y Mega con la turca en 25. Con razón estos hijos de puta salen electos una y otra vez. .. Infórmese!	910

Fuente: Elaboración Propia.

La tabla 4.6 muestra los cinco mensajes con más retweeteos del set de datos. Se puede apreciar que cada uno de ellos habla de un tema diferente y fue realizado por un usuario diferente.

Número de Retweets y Favoritos por usuario

Siendo el objetivo poder determinar la influencia de un usuario, se sumó la cantidad de retweets y favoritos que tiene cada usuario del set analizado obteniéndose los valores agregados.

Tabla 4.7: Cuentas con mayor cantidad de RT

Cuenta	05/01 al 11/01	12/01 al 18/01	22/12 al 28/12	29/12 al 04/01	Total
biobio	5975	6935	10212	10463	33585
Cooperativa	4737	6024	8380	8577	27718
latercera	3346	5400	6930	6334	22010
T13	2115	2787	4209	3749	12860
soychilecl	2268	3197	3697	3634	12796

Fuente: Elaboración Propia.

La tabla 4.7 muestra las cuentas con mayor cantidad de *retweets*, sin mucha sorpresa, son todas de noticias, lo refuerza la idea sobre el rol que tiene Twitter como un medio fuertemente informativo[24]. Hay que destacar que estas cuentas no son las mismas que aquellas que tienen un mayor número de mensajes realizados en la red social, pero, sin lugar a dudas, son cuentas que realizan una gran cantidad de mensajes.

Número de Retweets por número de tweets

El número de RT para una cuenta puede ser siempre alto si se encuentra escribiendo mensajes continuamente, como el caso de los perfiles de noticias vistos anteriormente, es por ello que se dividen la cantidad de retweets por el número de mensajes que genera la cuenta para tener una visión del *promedio* de retweets que tiene un perfil y así comparar.

Tabla 4.8: Cuentas con mayor cantidad de RT por número de tweets

Cuenta	05/01 al 11/01	12/01 al 18/01	22/12 al 28/12	29/12 al 04/01	Total
EnciclopediaCL	1438.5	743.65	861.75	744.33	3788.23
C1audioBravo	0	0	3298	0	3298
malaimagen	3	44.08	2075.5	305.67	2428.25
JorgeVilchesV	59.76	520.3	743.8	1047.83	2371.69
LaTiaEvelyn	624.97	261.67	698.25	661.27	2246.16

Fuente: Elaboración Propia.

La tabla 4.8 muestra este ratio por cada semana ordenando las cuentas por el total obtenido en los 28 días. Tres cuentas de humor se mantienen en el top cinco de este promedio, mientras que el perfil ClaudioBravo con solo un mensaje logra una gran cantidad de retweets y la cuenta JorgeVilchesV es de un usuario que solo realiza *spam* en la red social.

Tabla 4.9: Primeros 5 usuarios

Perfil	ScoreRT	Suma RT	N° Tweets	Score TS-LDA	Puntaje
DonosoPavez	30.07	934	6	3.53	106.05
sebastianpinera	27.1	394	1	3.59	97.22
ciper	17.38	821	12	2.89	50.21
TaniaMelnick	21.28	1110	14	1.68	35.69
JorgeAlis	68.07	2855	10	0.52	35.2

Fuente: Elaboración propia.

4.5.4. Perfiles influyentes

Finalizando el capítulo se tiene la influencia estimada de distintas personas dado la propagación que generan y lo interesante de los tópicos en los que hablan. La separación de temas se realizó por aquellos con una gran cantidad de tópicos asociados, como se vio en la figura 4.8, estos son política y deportes. No se utilza los tópicos relacionados con *cháchara* porque suelen tener puntajes muy bajos, al igual que *saludos*, por su parte noticias está presente en muchos tópicos porque Twitter es una fuente en general de información y como se vió en la tabla 4.7, la mayoría de las cuentas que tienen mayor cantidad de RT son de noticias.

No se utilizaron más tópicos por la baja cantidad de temas relacionados con los que cuentan, este problema es tratado más adelante.

General

Dado lo visto durante este capítulo, se puede obtener de manera final un ranking de perfiles que crean tweets interesantes que tienen repercusión en la red social, en otras palabras que son *influyentes*. La manera en que se representa la repercusión en la red provoca diferencias en el orden de los perfiles. Tomando en consideración la ecuación de se tiene que los primeros perfiles no necesariamente son aquellos con más retweets o mayor ratio de retweets por mensajes generados.

Observando la tabla 4.9 se aprecia que inclusive la cuenta *sebastianpinera* tiene un puntaje elevado a pesar de solo poseer un tweet. Por su parte, la cuenta *JorgeAlis*, que es de un comediante, es la única de las cuentas que tiene un puntaje bajo 1 en Score TS-LDA normalizado.

Política

En el mundo de la política se asociaron 14 tópicos, los cuales gracias al puntaje asignado y la distribución $p(t|u)$ de cada usuario se asignaron valores de lo *interesante* asociado a su participación en Twitter.

Tabla 4.10: Primeros 10 usuarios tema política

Perfil	ScoreRT	Suma RT	N° Tweets	Score TS-LDA	Puntaje
DonosoPavez	30.07	934	6	5.98	179.76
joseantoniokast	35.15	4501	50	2.53	89
JorgeAlis	68.07	2855	10	1.21	82.36
ciper	17.38	821	12	4.63	80.5
allamand	16.05	403	4	4.08	65.5
ja_richards	11.21	1657	60	5.84	65.45
melissasepulvda	7.15	105	1	7.8	55.75
EnciclopediaCL	71.83	9186	50	0.74	52.94
LaTiaEvelyn	37.09	7580	92	1.4	51.94
ChaoGirardi	4.5	84	2	10.95	49.23

Fuente: Elaboración propia.

Tabla 4.11: Primeros 10 usuarios tema deportes

Perfil	ScoreRT	Suma RT	N° Tweets	Score TS-LDA	Puntaje
IgnacioCasale	13.5	706	14	4.94	66.75
Heinekencl	13.47	864	19	2.2	29.59
EnciclopediaCL	71.83	9186	50	0.31	22.56
DonosoPavez	30.07	934	6	0.74	22.28
CerveceriaKross	4.82	180	8	3.69	17.82
maggifaundez	11.28	208	2	1.55	17.43
PeugeotCL	1.81	195	37	9.38	16.99
ColoColo	4.64	2817	346	3.47	16.1
malaimagen	51.4	5364	38	0.27	14.11
tvMOTOTEMATICOS	1.01	115	38	12.37	12.45

Fuente: Elaboración propia.

La tabla 4.10 muestra a los primeros 10 perfiles donde se encuentran políticos conocidos y algunas cuentas de comedia que tienen una gran cantidad de RT, lo que provoca que aparezcan en el ranking.

Deportes

Por el lado de los deportes se asociarion una cantidad de 11 de tópicos. Cabe destacar que esto no es necesariamente fútbol, sino que deporte en general.

Como se observa en la tabla 4.11, la mayoría de los perfiles son de deportistas y marcas que apoyan a los deportistas nacionales. Dado que en Chile se realiza desde el 2009 el Rally Dakar[8], aparece en el primer lugar un piloto de rally. También se observa que las cuentas que

[8]El Rally Dakar es una competencia anual de *rally raid* donde participan automóviles, camiones, moto-

tienen un *Score TS-LDA* menor a 1 no se asocian a primera vista al deporte, pero puede que hayan hecho alguna mención a los deportistas, por lo que su puntaje está arriba de la media.

cicletas y cuadriciclos. Su nombre radica en la meta del *rally*, que solía ser la ciudad de Dakar en Senegal. Dado un cambio de ubicación por temas de seguridad, desde el 2009 se desarrolla en países sudamericanos donde Chile ha estado presente. Para más información visite: `http://www.dakar.com`

Capítulo 5

Predicción de Influencia

Este capítulo tratará de inferir solo con los datos existentes en twitter el score de influencia dado en el capítulo anterior. Para ello se pretende ver que variables de las obtenibles por la API de Twitter se correlacionan con el score de influencia (más allá de las variables del *score* mismo). Además, se pretende ver si estos datos al cambiar en el tiempo (como la cantidad de retweets que reciben los tweets de un usuario) son indicadores para predecir la influencia.

La predicción no ha dejado de ser un tema controversial para distintos investigadores que no concuerdan sobre su validez, aun así es usada en diversos ámbitos para poder planificar sobre sucesos futuros[43]. En el caso de Twitter, predecir la influencia llevaría a poder adelantarse a quiénes serán más escuchados y así tomar esos perfiles en consideración con anticipación para distintos fines.

Dado que, como se vió en el capítulo anterior, se tomó en consideración la repercusión de un perfil en las redes sociales como la generación de contenido interesante y la cantidad de retweets que tienen un perfil por cantidad de tweets, la influencia se puede descomponer en dos valores a predecir: el *score* de interés para un tema en particular y la cantidad de retweets que tiene el perfil por número de tweets.

Para poder predecir estos valores como series temporales se ha de realizar el supuesto en que las variables cumplen la hipótesis de recursividad temporal, es decir las variables no dependen de los eventos futuros, y que el proceso ha de ser invertible, es decir que las variables pasadas van siendo menos influyentes sobre las variables futuras. Las predicciones fueron realizadas en R[64] con el paquete *forecast*[33].

Los modelos más usados para predecir suelen ser ARIMA y ANN, es por ello que se utilizaron ambos modelos para evaluar la predicción a realizar en este trabajo. Ambos modelos se comportan de manera parecida cuando existe una gran cantidad de observaciones y ANN precide de mejor manera cuando existe una menor cantidad de datos[74]. La cantidad mínima de datos para obtener un modelo confiable difiere de aplicación a aplicación, y por ello se recomienda usar "tanta data como sea posible"[34]. De todos modos, la cantidad de observaciones con las que se debe contar está muy relacionada con la variabilidad de los datos; es decir, entre más variables sean, más observaciones se necesitan.

5.1. Predicción con ARIMA

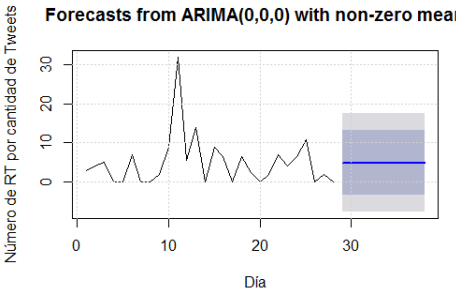
Forecast cuenta con un módulo de elección automática del mejor modelo de autorregresión ARIMA en base a su valor AIC, AICc o BIC, es por ello que se utiliza este comando para la predicción de la cantidad de retweets por número de tweets. La temporalidad de los datos usados es de un día, por lo que si se quiere predecir para cada día siguiente se toma la frecuencia como 1. Ahora bien, como una persona en Twitter no siempre estará produciendo mensajes con retweet cada día se toma como unidad de tiempo una semana y con ello la frecuencia pasa a ser de 7.

Para poder evaluar el modelo predictivo se utilizó una cuenta que tuviera participación en la red durante cada una de las cuatro semanas en las cuales fueron tomados los datos. Eligiendo una cuenta al azar, se utilizó el perfil de *huichalaf* para ver si existen patrones observables en el mes de datos recolectados.

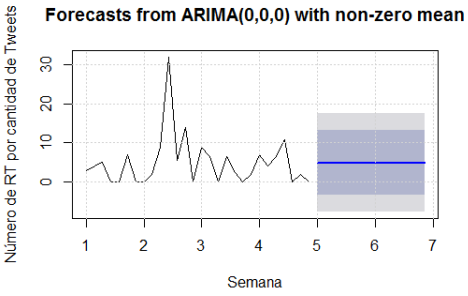
La figura 5.1 muestra resultados de la predicción con ARIMA, realizada con *auto.arima*, siendo los intervalos de predicción al 80 % en gris oscuro y al 95 % con gris claro. La línea azul muestra el resultado medio de la predicción, que es constante e igual para ambos tipos de frecuencia siendo este valor un poco menor que 5.

Figura 5.1: Resultados de la predicción ARIMA

(a) Frecuencia diaria



(b) Frecuencia semanal



Fuente: Elaboración propia.

Los resultados obtenidos son iguales al promedio mensual que generó este perfil en el lapso de datos recolectados, lo que no agrega nueva información acerca de los posibles valores futuros. De todos modos, esto siempre ha de ocurrir cuando el valor de d $= 0$ y $\phi_0 \neq 0$. La cantidad de datos es tan pequeña que el mejor modelo para predecir solo dará la media de los datos. Dado que redes neuronales se realiza mejores predicciones que los modelos ARIMA cuando se tiene una pequeña cantidad de datos [74], se procedió a realizar las predicciones con redes neuronales.

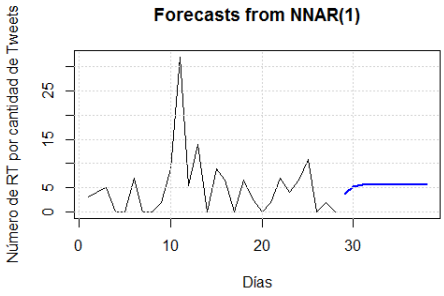
5.2. Predicción con Redes Neuronales Artificiales

Las redes neuronales aritificiales (ANN, por sus siglas en inglés) cuentan con gran popularidad para la predicción de datos, como también para modelos de clasificación, ya que son flexibles y confiables. El módulo *nnetar* de *forecast* usa redes neuronales con una capa oculta para la predicción de series temporales.

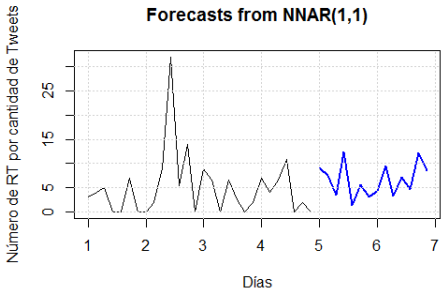
La figura 5.2 muestra los resultados de la predicción con ANN, en esta ocasión el paquete *forecast* no cuenta con intervalos de predicción para redes neuronales, por lo que solo se muestra la línea promedio de la predicción. A diferencia de ARIMA, en este caso se tiene una tendencia menos lineal cuando se utiliza una frecuencia de 7 para la unidad de tiempo, pero aun así los valores promedios semanales predecidos, 6,15 y 6,96 son muy cercanos al promedio de la muestra.

Figura 5.2: Resultados de la predicción ANN

(a) Frecuencia diaria



(b) Frecuencia semanal



Fuente: Elaboración propia.

Aunque teóricamente uno puede predecir con un pequeño número de datos, es la variabilidad de los datos lo que obliga a contar con una mayor cantidad de información previa[34], es por esto, juntos a los resultados de ANN y ARIMA, que no se considera los resultados como confiables para la predicción del número de retweets por cantidad de tweets, y por consecuencia la predictibilidad de $p(t|u)$ por cada día o semana se ve comprometida de la misma manera al tener incluso una periocidad menor.

Capítulo 6

Evaluación del Modelo y Discusión

Cuando se realizan estudios suelen surgir ciertas inquietudes acerca de que pasaría si se movieran tales variables. En este capítulo se tratará acerca de una evaluación del modelo propuesto y algunos análisis de sensibilidad para observar la consecuencia en los factores propuestos.

6.1. Resultados obtenidos y análisis de sensibilidad

6.1.1. LDA

Los parámetros para afinar el modelo LDA son α, β y el número de tópicos. Dado que α y β representan que tanto peso se le da a la asignación de tópicos[27], estos valores no fueron modificados. Como se vió en el capítulo 4, el número de tópicos asignados fue de 200, de todos modos se realizó LDA con 500 y 100 tópicos. Estos valores se encuentran en el apéndice A.

Con una asignación manual del etiquetado de tópicos se determinó a qué corresponde cada uno. En la figura 4.8 se tiene que la asigación de tópicos los clasifica mayormente en cháchara, por lo que es difícil poder hacer una segmentación detallada de temas de interés pada poder ver a los usuarios influyentes de ellos. Eso es algo en contra a la metodología propuesta, ya que resulta improbable poder distinguir en temas como minería o medio ambiente por la nula presencia de estos en el universo de tópicos obtenidos tal como se vió en el subcapítulo 4.5.1. Estos temas tendrían que ser parte de noticias actuales para que estén más presentes en los resultados de LDA.

En cuanto al análisis de sensibilidad, se tiene que una de las mayores implicancias de aumentar el número de tópicos en esta metodología radica en la separación de dos temas que podrían ser considerados solo uno, por ejemplo en la tabla 6.1 se puede observar que con 200 y 500 tópicos se separan en dos el tópico de la sub-tabla *a*). Dado que se quiere obtener temas latentes en el tiempo no se utilizó un número pequeño de tópicos para realizar el modelo, además como se mencionó anteriormente existen problemas en la cantidad de tópicos útiles

Tabla 6.1: Tópicos *outliers*

a) 100

Tópico	P. 1	P. 2	P. 3	P. 4	P. 5	Puntaje W
63	vin	gtgt	renac	arriend	mar	10.36

b) 200

Tópico	P. 1	P. 2	P. 3	P. 4	P. 5	Puntaje W
93	person	excelent	vin	vist	info	22.98
78	vin	mar	diari	jardin	arriend	14.72

c) 500

Tópico	P. 1	P. 2	P. 3	P. 4	P. 5	Puntaje W
364	vin	mar	diari	jardin	gtgt	18.79
379	person	excelent	vist	vin	info	18.35

Fuente: Elaboración propia.

para algún análisis, lo que sugiere que se debió trabajar más tópicos.

6.1.2. TS-LDA

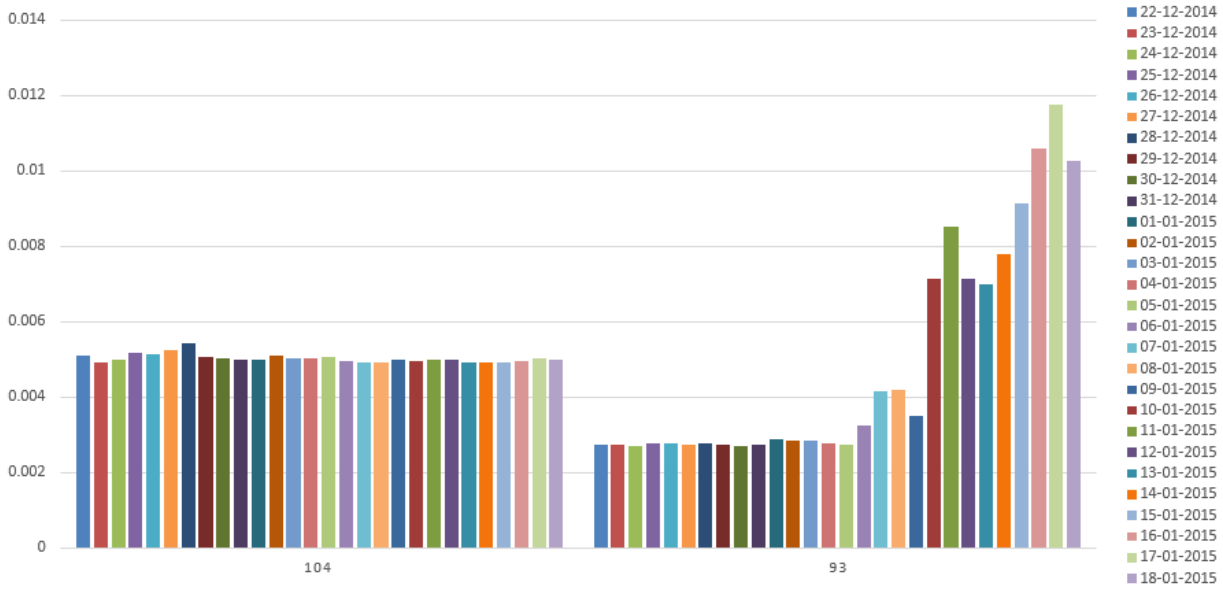
En cuanto a TS-LDA tenemos que el modelo presentado muestra un problema con las cuentas que ofrecen arriendos en la costa. Este es el caso, cuando se tienen 200 tópicos, de los temas n° 78 y 93; repitiendose en mismo problema cuando se ejecutan más o menos tópicos. En la tabla 6.1 se puede observar como los valores con un mayor puntaje W (el puntaje de interés) son acerca de arriendos en Viña del Mar. Estos tópicos cuentan con una entropía espacial y temporal bastante baja, lo que provoca que parezcan ser de un interés repentino.

La figura 6.1 muestra como los tópicos con más y menos entropía espacial se distribuyen entre las fechas, es decir $p(t|s)$. Esta probabilidad no es usada por las fórmulas de la sección 4.4.2, pero es una consecuencia de $p(s|t)$ como se ve en la ecuación 6.1.1 y sirve para ilustrar como un tópico se mueve en el tiempo.

$$p(s|t) = p(s) * p(t|s)/p(t) \tag{6.1.1}$$

Así, tenemos que el tópico con mayor entropía espacial se distribuye de una manera pareja durante todos los días, lo que también afecta a su entropía temporal, por otro lado el tópico con menor entropía temporal apenas es mencionado en los primeros días lo que hace creer al modelo que se vuelve interesante dada la explosión de los últimos días. Lamentablemente al ser arriendos uno puede suponer que estarán presentes en toda la temporada veraniega, y por ello no deberían tener un puntaje más alto que, por ejemplo, el tópico 167. Este tópico, como se ve en la figura 6.2 tiene dos fechas donde es hablado bastante, el 30 y 31 de diciembre

Figura 6.1: Probabilidad del tópico dado una fecha para los tópicos 104 y 93 de 200



Fuente: Elaboración propia.

del 2014. Al observar la tabla A.2 se puede apreciar que el tópico trata de la ex ministra Helina Molina, donde por tales fechas realizó dichos sobre el aborto que la llevaron a salir del ministerio de Salud[25].

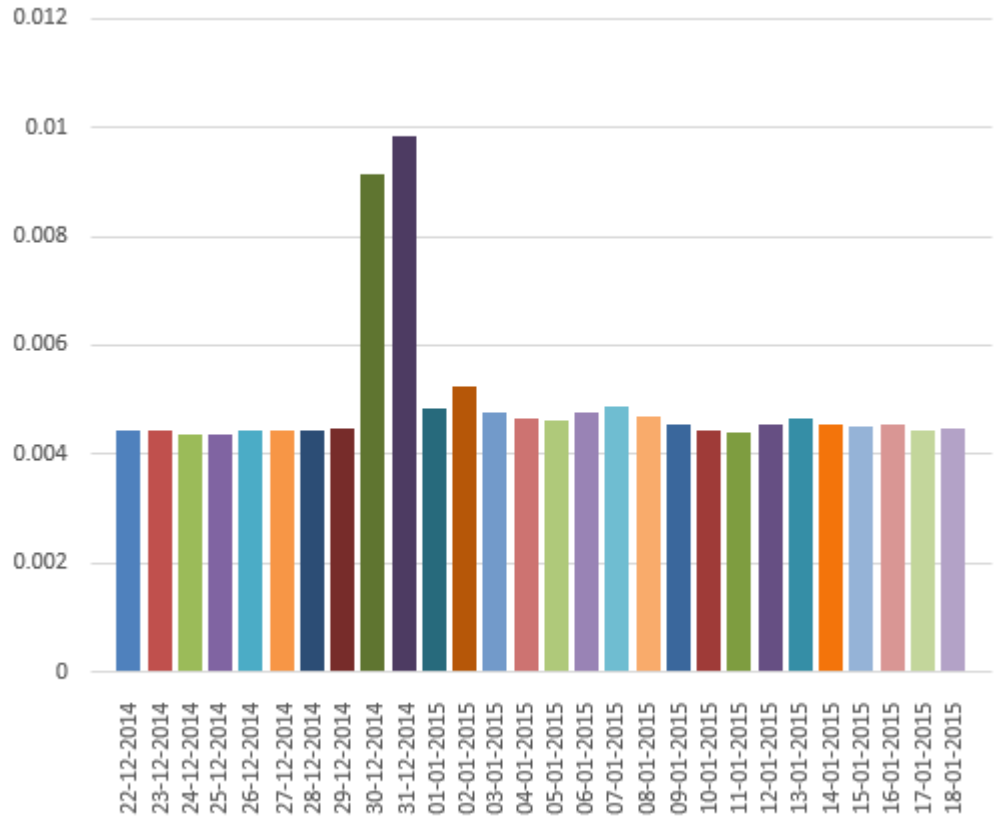
Con lo anterior podemos dar cuenta que TS-LDA es capaz de obtener temas del momento, pero que en esta aplicación falla por las cuentas que realizan *spam* continuo en la red. Es de importancia que se eliminen las cuentas que realizan *spam* durante el procesamiento de los datos, para evitar estos incovenientes.

Con la asignación de un puntaje 0 para los tópicos *outliers* se pudo subsanar en cierta medida el problema pero luego se tiene que el tópico n° 123 también es producto de una cuenta *spam* por lo que el problema persiste. Por su parte, estas cuentas no realizan de manera automática una gran cantidad de RT, lo que las lleva a no aparecer en los primeros lugares de los rankings de influencia.

6.1.3. Cuentas Influyentes

Para obtener las cuentas influyentes se tienen dos factores de importancia: La cantidad de retweets por número de tweets y el *score* de TS-LDA. Este último valor se normaliza para poder dejar con valores negativos aquellas cuentas que no hablan ni siquiera la mitad acerca del tema a tratar. Si no se realizara esta normalización se provocaría que algunos perfiles que tienen una gran cantidad de retweets por número de tweets aparecieran de los primeros lugares sin que se asocien al tema a tratar.

Figura 6.2: Probabilidad del tópico dado una fecha para el tópico 167 de 200



Fuente: Elaboración propia.

Tabla 6.2: Primeros 10 usuarios tema política sin normalizar

Perfil	ScoreRT	Suma RT	N° Tweets	Score TS-LDA	Puntaje
C1audioBravo	1811.61	3298	2	0.18	66.75
JorgeAlis	684.6	2855	10	0.26	29.59
EnciclopediaCL	722.36	9186	50	0.24	22.56
DonosoPavez	302.91	934	6	0.48	22.28
malaimagen	517.14	5364	38	0.23	17.82
joseantoniokast	353.94	4501	50	0.32	17.43
LaTiaEvelyn	373.45	7580	92	0.27	16.99
lafundacionsol	436.24	7710	77	0.22	16.1
ciper	175.49	821	12	0.42	14.11
YolandaSultanaH	376.22	12439	170	0.19	12.45

Fuente: Elaboración propia.

La tabla 6.2 muestra las consecuencias de no normalizar el puntaje de TS-LDA. En los temas de política aparecen cuentas que tienen un puntaje muy pequeño pero dado al alto ScoreRT que tienen aprecen en la lista, como el caso de *C1audioBravo* o *YolandaSultanaH*. Ambas cuentas no se asocian con el tema de política.

Para poder resolver este inconveniente también es posible pedir que las cuentas a considerar pasen de cierto umbral en su puntaje TS-LDA, pero esto llevaría a realizar un análisis de histograma en cada iteración de modo de poder saber que umbral utilizar para tener el tanto por ciento de los usuarios que hablan de cierto tema.

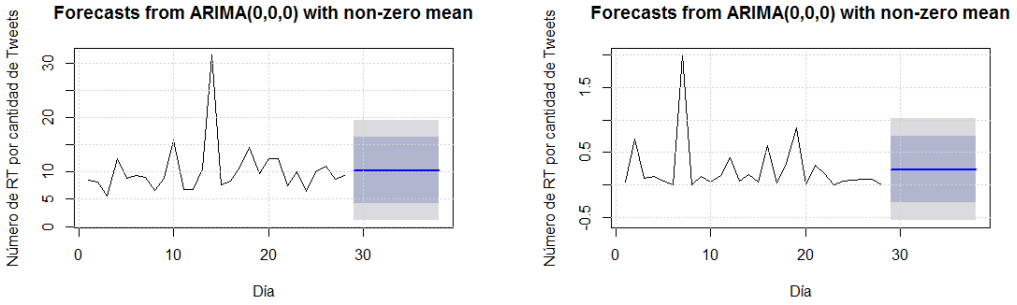
6.1.4. Modelos Predictivos

Los modelos predictivos vistos en el capítulo 5 no llevan a una predicción confiable. La capacidad de los algoritmos para poder predecir el número de retweets por número de tweets no es confiable para la cantidad de datos. Eso sí, esto se realizó para solo una cuenta al azar dentro del set de datos que tuviera participación todas las semanas, por lo que cabe preguntar si es un comportamiento común en la data.

En las figuras 6.3 se observan las predicciones en ARIMA para las cuentas de *latercera* y *TecnoFury*, donde se puede apreciar el mismo patrón que lo visto en el capítulo 5. La predicción con cualquier cuenta presenta igual comportamiento, donde no es posible predecir semanalmente la cantidad de retweets por número de tweets, obteniéndose apróximadamente solo el promedio de los datos.

Figura 6.3: Resultados de la predicción ARIMA para las cuentas *latercera* y *TecnoFury*

(a) *latercera* (b) *TecnoFury*



Fuente: Elaboración propia.

Por lo tanto se ha de requerir una mayor cantidad de información antes de realizar las predicciones de series temporales. No es posible, con la data que hay, poder observar puntos de inflexión o tendencias en el tiempo para inferir el impacto que tendrá un perfil en un futuro cercano.

Los resultados obtenidos por *auto.arima* son del modelo ARIMA más simple, un ARI-

MA(0,0,0), en cual dará siempre la media de los datos como predicción. Aunque se pueden realizar análisis con otros grados del modelo, *auto.arima* elige el mejor modelo en base al valor AIC, AICc o BIC. Además, algunos autores recomiendan una cantidad de 60 datos para poder predecir[28], mientras que otros indican que depende la variabilidad de los datos[34]. Estos datos son demasiado variables, por lo que se vuelve a recomendar una recopilación más continua y por un mayor periodo de tiempo para poder realizar alguna predicción.

6.2. Discusión

La discusión se centrará brevemente acerca del cumplimiento de los objetivos específicos planteados, para luego mencionar algunos problemas que son parte del no cumplieminto de ciertos objetivos.

6.2.1. Sobre los resultados

Para ver si los resultados cumplieron los objetivos de la memoria, se analiza cada objetivo específico propuesto al comienzo de este trabajo y se comenta en base a lo realizado durante el semestre.

Los objetivos especificos de la memoria eran los siguientes:

- *Establecer el estado del arte de las técnicas relacionadas con índices de medición de influencia de perfiles en Twitter*
  Este objetivo se pudo cumplir con la última parte del capítulo 2 junto al capítulo 3, donde se presentaron los diversos modelos que existen para medir influencia en Twitter.
- *Definir un modelo a utilizar que permita obtener un indicador de medición del impacto de perfiles en twitter*
  Este objetivo fue cumplido al finalizar el capítulo 3, donde se definió que el modelo a utilizar sería TS-LDA dado que permite darle un score de interés de los diversos tópicos, y con ello junto a los retweets se puede obtener un indicador del impacto del perfil.
- *En base al modelo elegido, crear un algoritmo que permita obtener un score de influencia de perfiles en Twitter de acuerdo a tópicos de interés y, si fuese posible y necesario, agregar velocidad de propagación*
  En la sección 4.1 se modeló un score de influencia de acuerdo a la capacidad de crear contenido interesante y propagarlo por la red social, pero como se vió más adelante, los tópicos de interés no fueron de fácil extracción y por lo tanto no se pudo segmentar entre los diversos temas que existen, por lo que este objetivo no fue cumplido en su totalidad. La velocidad de propagación de un tweet no fue tomada en cuenta dado que se subestimó el tiempo que se invertió en el trabajo, por lo que no fue posible agregar tal varibale.
- *Crear un modelo que entregue un grupo de perfiles en Twitter que tengan alta probabilidad de ser influyentes en un futuro próximo*
  En el capítulo 5 se vieron los resultados que se presentan al modelar series de tiempo

para la cantidad de retweets por número de tweets que realiza un perfil, llegándose a la conclusión que la cantidad de datos es muy baja para realizar predicciones confiables, es por ello que este objetivo no pudo ser cumplido completamente, sin embargo, se estudian los motivos que impiden mejorar los resultados y se realizan propuestas a futuro.

- *Evaluar y concluir en base a los resultados obtenidos*
  Este objetivo se cumple en el capítulo actual y en el siguiente.

6.2.2. Problemas detectados

Durante la realización de la memoria se pudieron apreciar ciertos problemas que llevaron al no cumplimiento de algunos objetivos vistos en la sección anterior.

- Pocos temas identificados
  Con la realización de LDA de 200 tópicos se pudieron identificar temas que contenían muy pocos tópicos lo que provocó que no fuera posible segmentar en varios grupos los perfiles recolectados.
- Outliers en los resultados de TS-LDA
  Los tópicos que obtuvieron un puntaje que se escapaba demasiado a los demás tópicos, provocan una distorsión en los resultados tal como se mencionó en la sección 6.1.2.
- Predicción poco confiable
  Los resultados del capítulo 5 mostraron que con una poca cantidad de datos la predicción se transforma solamente en inferir la media de la data. Durante este trabajo existió una tardía implementación del sistema de recolección de Tweets, pudiéndose recién poner en marcha a mediados de diciembre del 2014, por lo que no fue posible obtener una mayor cantidad de información y de manera temprana.

Dado los resultados, se puede apreciar que hay un largo camino por delante para poder predecir de una manera más certera usuarios que serán influyentes en la red social, sin embargo, tener una lista de usuarios influyentes puede ser de utilidad para diversas funciones y servicios que sean de interés. Esto quiere decir que, al contar con un *ranking* de importancia de los perfiles de una red se puede priorizar a que usuarios tomar en cuenta para, por ejemplo, realizar campañas de marketing que luego repercutan en la red de boca en boca[65]. Otras utilidades de tener una priorización de los usuarios son vistas en el capítulo siguiente.

Finalmente se tiene que, los problemas discutidos a raíz de los resultados y los análisis llevan a sugerir distintas propuestas para sobrellevar las dificultas listadas con anterioridad que son presentadas en el capítulo siguiente como un posible trabajo a futuro.

Capítulo 7

Trabajo Futuro y Conclusiones

Al final de cada trabajo es importante realizar discusiones acerca de los resultados asociados, las dificultades del proceso, qué se puede mejorar y qué desafíos quedan por delante. Este capítulo final tratará acerca de un trabajo futuro propuesto, como también de las conclusiones que se desprenden de la realización de la memoria.

7.1. Trabajo futuro

En esta sección se verá en una primera instancia las posibles mejoras al trabajo realizado para resolver los problemas vistos en la sección anterior y luego se verán algunas aplicaciones y futuras investigaciones que se pueden desprender del trabajo realizado.

7.1.1. Mejoras al modelo de definición de influencia

El vista al trabajo realizado es posible poder agregar diversas variables que puedan mejorar el método de caracterización de influencia presentado en este trabajo. Algunos de los factores a agregar podrían ser

Aplicar LDA a gran escala

Ligado al primer problema, una solución es aumentar la cantidad de datos o la cantidad de tópicos con los cuales se realiza LDA. Dado que la aplicación de Latent Dirichlet Allocation para la identificación de tópicos y del modelo TS-LDA conlleva una gran cantidad de tiempo. Si se requiere utilizar este modelo para identificar los tweets interesante o los tópicos hablados en una data muy grande sería aconsejable utilizar algún otro enfoque que aproveche de mejor manera los recursos o utilice algún modelo de programación paralela como MapReduce.

MapReduce[16] es un modelo de programación e implementación asociada para procesar

largos set de datos con un algoritmo paralelo y distribuido en un cluster, siendo en un comienzo utilizado por Google ha ido ganando adeptos y ha sido adaptado con librerias en diversoso lenguajes. Apache Hadoop[1] cuenta con una implementación open source de esto modelo de programación.

Dado que MapReduce puede distribuir la carga de trabajo en distintos nodos existen distintas implementaciones de LDA realizadas en MapReduce, como por ejemplo Mr. LDA[88], PDLA[83], o inclusive el proyecto Mahout[2] de Apache que cuenta con un módulo de LDA. Durante este trabajo no se realizó ningún análisis sobre estas implementaciones para poder determinar cual sería idónea para aplicar TS-LDA, de todos modos los creadores de PDLA realizaron una mejora a su modelo llamda PDLA+[42] el cual presenta un rendimiento casi lineal hasta cerca de 400 computadores (es decir, 2 computadores es el doble de rápido y consecutivamente), por lo cual es bastante útil para la división de trabajo, por su parte los autores de Mr. LDA durante su elaboración dicen superar a Mahout en rendimiento y velocidad, aunque dado que Mahout es un proyecto libre que es mejorado constantemente puede que haya mejorado.

Disminuir la memoria utilizada

Para poder contener los datos de un poco más de 2,000,000 de tweets y asignarlos a 500 tópicos se tuvo que ocupar alrededor de 40gb de RAM. Esto es una gran cantidad de memoria utilizada, por lo que si bien es posible paralelizar el proceso de LDA como se mencionó en el punto anterior, también es factible cambiar el tipo de variables a utilizar.

Jgibblda utiliza las varibales *double* en JAVA, lo que ocupa una cantidad de 8 bytes. Por su parte, las variables *float* en JAVA utilizan la mitad de memoria, solo 4 bytes, si bien no tienen la misma presición de una variable *double*, en el modelo de LDA puede que no sea de gran problema la menor asginación de decimales de un *float* por sobre una disminución de la memoria utilizada. Aun cuando un *double* ocupa el doble de espacio que un tipo *float*, el rendimiento de procesamiento no es el doble de lento[29], por lo que el rendimiento del procesador no debería ser un factor tan determinante a considerar.

Utilización de un diccionario más completo

Ya más ligado al segundo problema detectado se tiene que se puede mejorar el uso del diccionario. El diccionario utilizado en el modelo fue el LIFACH, el cual como se mencionó es un repositorio bastante completo de las palabras usadas del español de Chile, pero estas se encuentran asociadas a su lema. En el trabajo, para llevar las palabaras a su raíz se utilizó el algoritmo de Porter (y también se aplico al diccionario), necesariamente no se da que queden las mismas palabras, por ejemplo, en el caso del LIFACH los diversos articulos definidos (el, la, lo) fueron agrupados a un solo artículo, provocándose que existen diferencias en las palabras usadas. El LIFACH no deja de ser un excelente recurso para poder determinar la frecuencia de las palabras de distintos medios, pero para la integridad de TS-LDA se pueden

[1] http://hadoop.apache.org/

Figura 7.1: Tres de los quinientos tópicos de LDA del 22/12/14 al 18/01/15

(a) Tópico 24

Palabra	Probabilidad
jajaj	0.057684402
obvi	0.051903176
po	0.042020637
jajajaj	0.041579616
xd	0.019009904
jajajajaj	0.014486534
pos	0.004892104
sip	0.004005549
ajjaj	0.001127971
pfff	0.001070718
jajajajajaj	9.99E-04
oye	9.80E-04
igual	9.53E-04
nop	5.65E-04
aaah	5.54E-04

(b) Tópico 26

Palabra	Probabilidad
xd	0.132711611
jajajaj	0.050372906
jajajajaj	0.018050685
hahah	0.01125166
hah	0.004524515
eso	0.003331156
oo	0.003168097
igual	0.002843611
lol	0.002462646
jajajajajajaj	0.001754148
jajaj	1.56E-03
sali	1.44E-03
ando	1.29E-03
jajajajjajaj	1.16E-03
hahahah	1.13E-03

(c) Tópico 38

Palabra	Probabilidad
jajajaj	0.122144483
xd	0.067386115
igual	0.01879257
jajajajajaj	0.013758295
chistos	0.011580689
rei	0.004576234
jajajajaj	0.002436145
siii	0.00177147
sip	0.001414436
jajajajajajaj	0.001180978
bkn	7.38E-04
jajajajajaaj	7.02E-04
uta	3.75E-04
terribl	3.74E-04
gracios	3.67E-04

Fuente: Elaboración propia.

obtener mayores extremos en los puntajes dando valor a una mayor cantidad de palabras al tener un diccionario más completo y que al aplicársele el algoritmo de Porter tenga mayor concordancia con los tokens que se van produciendo.

Lematización de palabras que producen ruido

¿Cuál es la diferencia entre *jajajaja*, *jajajajajjajajajaj* o *jajkasjdkajsdkajkasjd*? Probablemente no mucha, todos estos tokens representan distintas formas de risa y aunque sea bastante complicado poder tener una expresión regular[2] para todas las formas de risa, se puede agrupar ciertos tipos (como por ejemplo tokens que contengan solo j y a) para que tengan un lema en común, así se podrían evitar ciertos ruidos en la aplicación de LDA.

En la figura 7.1 se puede apreciar como tres tópicos presentan diversas formas de risa. Aunque existen muchas más deformaciones de la risa, las más comunes suelen solo ocupar j y a, por lo se podrían reducir estos casos, donde incluso el tópico 26 muestra 6 formas diferentes de risa en las 15 palabras con más probabilidad de pertenecer a dicho tópico.

[2]Una expresión regular es una secuencia de caracteres para poder formar un patrón de búsqueda. Su uso está extendido en diversos lenguajes y, aunque no es lo más eficente, si es muy flexible por lo que su uso está bastante extendido.

Consideración de más variables

Dentro del trabajo se tomó un gran énfasis en poder reproducir los resultados de [86] para poder dilucidar lo interesante de un tweet dentro de la gran cantidad de tweets que se generan. Este enfoque deja de lado algunas consideraciones que se podrían haber utilizado, y que se pueden agregar, para poder mejorar la definición de influencia propuesta. Entre ellas se tienen:

- La velocidad de propagación de un tweet:
  En un principio era una posible idea a utilizar, la velocidad de propagación de un tweet que tiene un usuario podría dar otro enfoque acerca de como se propaga las acciones de una cuenta en Twitter.
- La forma de la red social:
  Es cierto que la estructura de la red social mirada desde *followers* y *followings* no representa una variable fuerte para agregar en la influencia, tal como se vio en los modelos del capítulo 3, pero se podría crear las aristas de la red social en base a la interación de dos personas a través de conversaciones que realizan en Twitter.
- La dispersión de los usuarios que hacen RT:
  Algunas cuentas en Twitter, como *JorgeVilchesV*, obtienen una gran cantidad de retweets que son realizados por los mismos perfiles, por lo que sería interesante ver el comportamiento de los retweets en función de la dispersión de las cuentas que los realizan.
- La inclusión de *hasgtags*:
  Aunque fueron eliminados por el uso que se les suele dar, los *hasgtags* son usados más seriamente por algunos usuarios, lo que llevaría a poder identificar el tema del que trata un tweet de mejor manera. Así, por ejemplo el tweet con más RT de la tabla 4.6 solo hace referencia a su tema en el *hasgtag* que utiliza, por lo que esta información fue eliminada durante pre procesamiento y sería de interés poder mantener esa relación al tema.

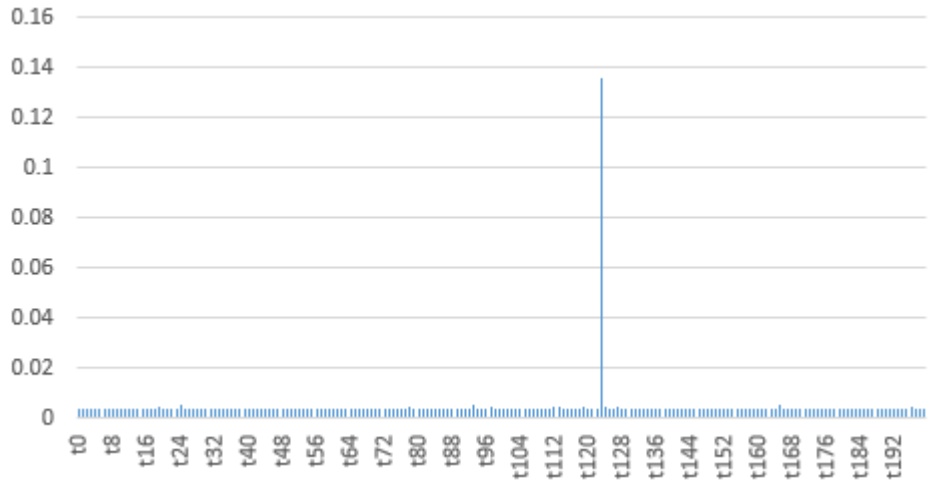
Detección de cuentas SPAM

Como se vió en la sección 6.1.2 las cuentas SPAM presentan un problema para el modelo actual, provocando que existan incoherencias con los resultados del modelo. No deja de ser cierto que los arriendos son un tema que aparece fuertemente en la temporada veraniega, pero es necesario evitar que pocas cuentas cuyo contenido se repite continuamente en la red distorsionen los modelos.

La figura 7.2 muestra la probabilidad de cada tópico para una cuenta *spam*. Se puede observar que al solo repetir información la probabilidad de los tópicos se centra en uno solo, que es el tópico que genera la distorsión.

Existen varias metodologías para detectar cuentas *spam* en Twitter[87, 6], pero ninguna utiliza un enfoque de análisis de tópicos, por lo cual podría ser una metodología a probar.

Figura 7.2: $p(t|u)$ para la cuenta *VinculoCL*



Fuente: Elaboración propia.

Etiquetado automático de tópicos en tweets

Según Yogesh Tewari y Rajesh Kawad[50] es posible etiquetar de manera automática los tweets que se van generando. Esta aproximación usa dos tópicos por tweet y se va generando de acuerdo a los tweets recibidos. Así, cada tweet es puesto en un archivo y luego leído por otro proceso que lo etiqueta según los tópicos que se generan. Aunque existe modelos que generan etiquetados de manera automática[69], el trabajo de Tewari utiliza varias de las herramientas que fueron usadas en este trabajo, como lo es Twitter4j o LDA, por lo que la generación automática de tópicos puede ser un tema a evaluar.

Recolección continua de datos

El tercer problema visto requiere que se guarde información durante un periodo más largo de tiempo. Si se toma en consideración que se necesitan entre 50 a 60 datos para poder predecir con confianza, se necestarían dos meses de datos. Si se toma en cuenta que la frecuencia debe ser de una semana, y que los datos deben ser acordes a la temporalidad deseada, se necesitaría un año de data de Twitter para poder realizar predicciones.

7.1.2. Segmentación a priori

En el primer problema detectado se mencionó la detección de pocos temas. Este problema puede ser resuelto con algunas de las mejoras vistas al modelo en la sección anterior, pero también se puede realizar el proceso con un enfoque distinto. Si en este trabajo se tomó todo el set de datos para poder aplicar los modelos propuestos, se puede primero segmentar para tener cierta población disponible para su análisis. Este enfoque podría ser de utilidad

de manera transversal para las empresas chilenas, al contar una población segmentada para los análisis de las redes sociales.

Determinación del género de los usuarios de Twitter

La segmentación de genero en Twitter puede ser un primer paso para tener conjuntos de gente diferenciados. Poder determinar el género de un perfil puede ayudar a categorizar de mejor manera la influencia que ejerce una cuenta dentro de un grupo de personas como también en la importancia que se le da sobre un tema a cierto grupo de opiniones. Así, por ejemplo, si se tienen temas relacionados con compras de zapatos de mujer, puede que sea más interesante ver a los usuarios influyentes dentro de la categoría del sexo femenino.

Aunque Twitter determina el género de sus usuarios, esta información no es pública por lo que se necesitaría de un segmentador propio. Un primer acercamiento a este problema puede ser utilizando la lista de nombres de hombres y mujeres que se registran año a año en el Registro Civil. Así, tomando los nombres de los usuarios en Twitter se puede clasificar si un token dado se encuentra en la lista de hombres o mujeres, para luego clasificar a esa persona dentro de un género. Claramente, se ha de tener en consideración una categoría indefinida, para poder recoger las cuentas cuyos nombres no aparecen en los litados y tambien los perfiles que representan tiendas o marcas.

Determinación del grupo socioeconómico de los usuarios de Twitter

La segmentación por grupo socieconómico ha sido usada largamente en Chile. Con tal información se podría ser más certero en la propagación de información en grupos de interés, es decir se podría ver el perfil de Twitter que genera más influencia en ciertos temas y que tenga además el nivel socieconómico deseado. Como una primera aproximación de realizar lo anterior se podría ver el nivel de escritura que presentan las personas.

Internacionalmente existen pocos estudios que vean el nivel de escritura y su relación con el nivel socio económico, pero dentro de la región podemos encontrar un estudio peruano[73] de una provincia en particular donde se muestra que existe una diferencia significativa entre el nivel socieconómico en el nivel de escritura, mas no así en la diferencia de género.

Para el caso chileno existen los resultados del SIMCE de escritura del año 2013[18], cuya publicación data del 2014, en el cual si bien se menciona que no existe una diferencia signficativa en el nivel socieconómico, existe una diferencia de 10 puntos porcentuales entre el estrato más bajo y el más alto que si es considerada grave por algunos expertos, como el representante de la Fundación Elige Educar, Joaquín Walker[77].

Es por ello que se podría tener una hipótesis donde “Las personas de un nivel socio económico más bajo escriben con un peor nivel lingüistico”, así se podría ver los comentarios en Twitter de las personas para poder segmentarlas en nivel socioeconómico bajo, alto e indefinido.

Es interesante notar además que, en la metodología utilizada en este trabajo, TS-LDA presenta en sus métricas la *Integridad* de un tópico, lo que muestra que tan aceptable (en relación a un diccionario de referencia) son las palabras usadas en un tópico en particular, y como se vió en 4.5.2 los resultados de la *Integridad* si reflejan, en cierta medida, que tan bien escrito está un tópico sobre otro por lo que la integridad de un perfil mostaría que este escribe mejor que otro con un menor puntaje.

7.1.3. Sistema de alertas prioritarias

Puede ser de interés de las empresas el poder contar con un sistema de alertas tempranas de enventos determinados. Esto se puede realizar programando un desencadenador que se active cuando un tweet cuente con determinadas características. Si se tienen tweets que van ingresando de un tema específico se puede generar la alerta cuando los mensajes comienzan a tener un tono negativo, así cuando un tema de interés empieza a tener mala reputación se pueden priorizar las alertas de acuerdo al grado de influencia de un perfil, de tal modo de no generar alertas cuando alguien con poca influencia escribe un mensaje muy negativo.

7.2. Conclusiones finales

La definición de influencia propuesta para los usuarios chilenos de Twitter permitió establecer cierto orden entre los perfiles recolectados. Si esta metodología es mejor o peor que otras no fue testeado ya que carece de uno de los puntos principales a estudiar: la segmentación por grupos de interés. La aplicación de LDA es potente, pero dada la naturaleza de Twitter, donde casi la mitad de sus tópicos se relacionan con conversaciones banales, fue difícil poder extraer tópicos para ser agrupados bajo un mismo tema de interés. Es por esto que se plantean varias soluciones para poder contrarrestar esta situación, soluciones que fueron apareciendo durante la realización de este trabajo de título pero que no fueron posibles de implementar durante el proceso.

Una de las mayores subestimaciones realizadas en esta memoria fue la capacidad de poder procesar la información en Twitter. Sin realizar muchos filtros al comienzo del trabajo, se obtuvieron alrededor de 50 millones de tweets, los cuales eran de un tardío procesamiento para un solo computador, sin mencionar el tiempo para realizar consultas de tipo `INSERT` o `SELECT` en una base de datos con tanta información. Así también, poder aplicar LDA con 500 tópicos a casi 2 millones de tweets requiere de una gran cantidad de memoria que guarde todos estos datos, además de una cantidad de tiempo que llega a ser más de 12 horas. El procesamiento de grandes volúmenes de información es uno de los retos de la llamada *Big Data*, y son resueltos en parte con ciertas metodologías que se podrían haber agregado al modelo que fueron mencionadas anteriormente. Nuevamente, este tipo de soluciones fueron descubiertas ya finalizando la memoria resultado imposible su implementación.

Siguiendo con los problemas, se tuvo que la predicción de variables requiere que los datos sean de un rango de temporalidad mayor. Los diversos inconvenientes que existieron para

poder poner en marcha el *streaming* de los datos llevaron a que se tuviera poca información para poder predecir. Aparte del *streaming*, se realizó una recolección de los últimos 3.200 tweets por usuario pero la diferencia temporal entre ambas recolecciones es demasiado grande como para pretender usarlas en conjunto para predecir. Es por ello que se ha de recolectar data con mucha anticipación si se pretender realizar predicciones para cualquier serie temporal de datos.

Por otro lado, las hipótesis planteadas no fueron comprobadas, lo que da a entender que el enfoque utilizado no es el más adecuado para enfrentar este problema. El ranking de influencia por temas da una idea acerca de la importancia relativa de un usuario a tal tópico de interés, pero la poca segmentación realizada juega en contra a la evaluación del modelo. También se tiene que para poder predecir se suelen necesitar unas 50 observaciones, lo que está lejos de las 28 observaciones usadas en este trabajo.

Finalmente, se tiene que el modelo de TS-LDA es bastante útil para determinar tópicos de interés y que estos pueden ser asignados a los usuarios, pero se recomienda realizar una segmentación *a priori* para poder diferenciar de mejor manera los grupos de interés y así encontrar los líderes de opinión, además de obtener la segmentación de Twitter que es anhelada por diversas empresas.

Bibliografía

[1] Popescu Adam. *Beyond Klout: Better Ways To Measure Social Media Influence*. 2013. URL: http://readwrite.com/2012/10/24/beyond-klout-better-ways-to-measure-social-media-influence (visitado 25-07-2014).

[2] Apache Software Foundation. *Apache Mahout:: Scalable machine-learning and data-mining library*. URL: http://mahout.apache.org.

[3] Ricardo Baeza-Yates, Berthier Ribeiro-Neto y col. *Modern information retrieval*. Vol. 463. ACM press New York, 1999.

[4] Eytan Bakshy y col. "Everyone's an Influencer: Quantifying Influence on Twitter". En: *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*. WSDM '11. Hong Kong, China: ACM, 2011, págs. 65-74. ISBN: 978-1-4503-0493-1. DOI: 10.1145/1935826.1935845. URL: http://doi.acm.org/10.1145/1935826.1935845.

[5] Parr Ben. *Klout Now Measures Your Influence on Facebook*. 2010. URL: http://mashable.com/2010/10/14/facebook-klout/ (visitado 25-07-2014).

[6] Fabricio Benevenuto y col. "Detecting spammers on twitter". En: *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*. Vol. 6. 2010, pág. 12.

[7] David M Blei y John D Lafferty. "Topic models". En: *Text mining: classification, clustering, and applications* 10 (2009), pág. 71.

[8] D.M. Blei, A.Y. Ng y M.I. Jordan. "Latent Dirichlet allocation". En: *Journal of Machine Learning Research* 3.4-5 (2003). cited By (since 1996)4800, págs. 993-1022. URL: http://www.scopus.com/inward/record.url?eid=2-s2.0-0141607824&partnerID=40&md5=505ce8839ae28d1cb56a7ff91bd0ad2d.

[9] Brandmetric. *Reportes Demo Brandmetric*. 2014. URL: http://demosbm.blogspot.com/ (visitado 29-11-2014).

[10] PEW Research Center. *Spring2013 Golabl Attitudes survey*. 2014. URL: http://www.pewglobal.org/2014/02/13/emerging-nations-embrace-internet-mobile-technology/ (visitado 29-08-2014).

[11] Meeyoung Cha y col. "Measuring User Influence in Twitter: The Million Follower Fallacy." En: *ICWSM* 10 (2010), págs. 10-17.

[12] CIPER. *Caso Penta: La caja negra de las platas políticas que sacude a la UDI*. 2015. URL: http://ciperchile.cl/2015/01/05/caso-penta-la-caja-negra-de-las-platas-politicas-que-sacude-a-la-udi/ (visitado 28-01-2015).

[13] ComScore. *Chile lidera en uso de redes sociales en Latinoamérica*. 2012. URL: http://noticias.universia.cl/en-portada/noticia/2012/11/22/983530/chile-lidera-uso-redes-sociales-latinoamerica.html (visitado 29-08-2014).

[14] Cooperativa. *Los tuiteros chilenos con más de 100 mil seguidores*. 2013. URL: http://www.cooperativa.cl/noticias/tecnologia/redes-sociales/twitter/los-tuiteros-chilenos-con-mas-de-100-mil-seguidores/2013-09-09/161246.html (visitado 29-08-2014).

[15] Rafael De Arce y Ramón Mahía. "Modelos Arima". En: *Departamento de Economía Aplicada. UDI Econometría e Informática. Universidad Autónoma de Madrid. Disponible en la World Wide Web: http://www.uam.es/personal_pdi/economicas/rarce/pdf/Box-jenkins.pdf http://db.doyma.es/cgibin/wdbcgi.exe/doyma/mrevista.fulltext* (2003).

[16] Jeffrey Dean y Sanjay Ghemawat. "MapReduce: simplified data processing on large clusters". En: *Communications of the ACM* 51.1 (2008), págs. 107-113.

[17] Pedro Domingos y Matt Richardson. "Mining the Network Value of Customers". En: *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '01. San Francisco, California: ACM, 2001, págs. 57-66. ISBN: 1-58113-391-X. DOI: 10.1145/502512.502525. URL: http://doi.acm.org/10.1145/502512.502525.

[18] Agencia de Calidad de la Educación. *Los aprendizajes en la escuela: Evaluación de Escritura*. 2014. URL: https://s3-us-west-2.amazonaws.com/resultados-simce-2013/Conferencia_prensa_resultados_Simce_Escritura_2013.pdf (visitado 28-01-2015).

[19] Roy Thomas Fielding. "Architectural styles and the design of network-based software architectures". Tesis doct. University of California, Irvine, 2000.

[20] William B Frakes y Ricardo Baeza-Yates. "Information retrieval: data structures and algorithms". En: (1992).

[21] Noah E Friedkin. *A structural theory of social influence*. Vol. 13. Cambridge University Press, 2006.

[22] Fundéu. *Hashtag: ¿Debería ponerse en cursiva? ¿Existe alguna traducción cuando hablamos de su uso en Twitter?* URL: http://www.fundeu.es/consulta/hashtag-30480/ (visitado 16-11-2014).

[23] Carlos Gershenson. "Artificial neural networks for beginners". En: *arXiv preprint cs/0308031* (2003).

[24] Julián González, Andrés Azócar y Andrés Scherman. *Encuesta de Caracterización de usuarios twitter en chile*. 2011. URL: http://www.slideshare.net/alestuardo/caracterizacin-de-usuarios-twitter-en-chile (visitado 29-08-2014).

[25] Tania González y Gonzalo Castillo. *Polémicos dichos sobre aborto sacan a Helia Molina del Ministerio de Salud, diarioUchile*. 2014. URL: http://radio.uchile.cl/2014/12/30/gobierno-acepta-renuncia-de-ministra-helia-molina (visitado 28-01-2015).

[26] Gregory Grefenstette. "Tokenization". En: *Syntactic Wordclass Tagging*. Springer, 1999, págs. 117-133.

[27] Thomas L Griffiths y Mark Steyvers. "Finding scientific topics". En: *Proceedings of the National Academy of Sciences* 101.suppl 1 (2004), págs. 5228-5235.

[28] James Douglas Hamilton. *Time series analysis*. Vol. 2. Princeton university press Princeton, 1994.

[29] Jeff Heaton. *Choosing Between Java's Float and Double*. 2011. URL: http://www.heatonresearch.com/content/choosing-between-java%E2%80%99s-float-and-double (visitado 29-08-2014).

[30] Liangjie Hong y Brian D. Davison. "Empirical Study of Topic Modeling in Twitter". En: *Proceedings of the First Workshop on Social Media Analytics*.

[31] Lan Huang. "A survey on web information retrieval technologies". En: *Computer Science Department, State University of New York at Stony Brook, NY* (2000), págs. 11794-4400.

[32] Rob J Hyndman y George Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2014.

[33] Rob J Hyndman, Yeasmin Khandakar y col. *Automatic time series for forecasting: the forecast package for R*. Inf. téc. Monash University, Department of Econometrics y Business Statistics, 2007.

[34] Rob J Hyndman, Andrey V Kostenko y col. "Minimum sample size requirements for seasonal forecasting models". En: *Foresight* 6.Spring (2007), págs. 12-15.

[35] Consultora IDC. *Chile es el líder en penetración de internet en Latinoamérica*. 2013. URL: http://www.cooperativa.cl/noticias/site/artic/20130529/asocfile/20130529145408/idc_barometro_2012_2h_chile_final.pdf (visitado 29-08-2014).

[36] JSON. *Introducing JSON*. 2014. URL: http://json.org/ (visitado 29-11-2014).

[37] Elihu Katz y Paul Felix Lazarsfeld. *Personal Influence, The part played by people in the flow of mass communications*. Transaction Publishers, 1970.

[38] Klout. *What are Klout Perks?* 2014. URL: https://klout.com/perks (visitado 25-07-2014).

[39] Mei Kobayashi y Koichi Takeda. "Information Retrieval on the Web". En: *ACM Comput. Surv.* 32.2 (jun. de 2000), págs. 144-173. ISSN: 0360-0300. DOI: 10.1145/358923.358934. URL: http://doi.acm.org/10.1145/358923.358934.

[40] Kred. *Kred Scoring Guide*. 2014. URL: http://kred.com/rules (visitado 25-07-2014).

[41] IBM y La Tercera. *Seguridad, educación y el futuro económico: los temas que más se tuitean en el país*. 2014. URL: http://www.latercera.com/noticia/tendencias/2014/10/659-600627-9-seguridad-educacion-y-el-futuro-economico-los-temas-que-mas-se-tuitean-en-el.shtml (visitado 10-04-2015).

[42] Zhiyuan Liu y col. "Plda+: Parallel latent dirichlet allocation with data placement and pipeline processing". En: *ACM Transactions on Intelligent Systems and Technology (TIST)* 2.3 (2011), pág. 26.

[43] Spyros Makridakis, Steven C Wheelwright y Rob J Hyndman. *Forecasting methods and applications*. John Wiley & Sons, 2008.

[44] Christopher D Manning, Prabhakar Raghavan e Hinrich Schütze. *Introduction to information retrieval*. Vol. 1. Cambridge university press Cambridge, 2008.

[45] Andrew Kachites McCallum. "MALLET: A Machine Learning for Language Toolkit". http://mallet.cs.umass.edu. 2002.

[46] Michelle McGinnis. *How Klout Is Measuring Your Influence on LinkedIn*. 2011. URL: http://blog.eloqua.com/klout-linkedin (visitado 25-07-2014).

[47] Geoffrey McLachlan y David Peel. *Finite mixture models*. Wiley-Interscience, 2000.

[48] Miller McPherson, Lynn Smith-Lovin y James M Cook. "Birds of a feather: Homophily in social networks". En: *Annual review of sociology* (2001), págs. 415-444. URL: http://www.jstor.org/stable/2678628.

[49] George Herbert Mead. "The genesis of the self and social control". En: *International journal of Ethics* (1925), págs. 251-277.

[50] Real-Time Topic Modeling of Microblogs. *Tewari, Yogesh and Kawad, Rajesh*. 2013. URL: http://www.oracle.com/technetwork/articles/java/micro-1925135.html (visitado 28-01-2015).

[51] Delia Mocanu y col. "The Twitter of Babel: Mapping World Languages through Microblogging Platforms". En: *PLoS ONE* 8.4 (abr. de 2013), e61981. DOI: 10.1371/

journal.pone.0061981. URL: http://dx.doi.org/10.1371%2Fjournal.pone.0061981.

[52] El Mundo.es. *Atentado Yihadista a Charlie Hebdo en Francia*. 2015. URL: http://www.elmundo.es/e/ch/charlie-hebdo.html (visitado 28-01-2015).

[53] Kamal Nigam y col. "Text Classification from Labeled and Unlabeled Documents using EM". English. En: *Machine Learning* 39.2-3 (2000), págs. 103-134. ISSN: 0885-6125. DOI: 10.1023/A:1007692713085. URL: http://dx.doi.org/10.1023/A%3A1007692713085.

[54] Oracle. *Java Language and Virtual Machine Specifications*. URL: http://docs.oracle.com/javase/specs/ (visitado 16-11-2014).

[55] Tim O'Reilly. *What Is Web 2.0*. 2005. URL: http://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html (visitado 29-08-2014).

[56] Lawrence Page y col. "The PageRank citation ranking: Bringing order to the web." En: (1999).

[57] Laurence AF Park y Kotagiri Ramamohanarao. "The sensitivity of latent dirichlet allocation for information retrieval". En: *Machine Learning and Knowledge Discovery in Databases*. Springer, 2009, págs. 176-188.

[58] Gillin Paul. *The Trouble with Klout*. 2011. URL: http://gillin.com/blog/2011/09/the-trouble-with-klout/ (visitado 25-07-2014).

[59] Xuan-Hieu Phan y Cam-Tu Nguyen. *Jgibblda: A java implementation of latent dirichlet allocation (lda) using gibbs sampling for parameter estimation and inference*. 2006.

[60] Addison Phillips y Mark Davis. *Tags for identifying languages*. Inf. téc. BCP 47, RFC 4646, September, 2006.

[61] Martin F Porter. "An algorithm for suffix stripping". En: *Program: electronic library and information systems* 14.3 (1980), págs. 130-137.

[62] PostgreSQL Core Team. *PostgreSQL: The world's most advanced open-source database*. The PostgreSQL Global Development Group. Vienna, Austria, 2015. URL: http://www.postgresql.org/.

[63] Emily Price. *Klout Makes Perks Easier to Claim*. 2012. URL: http://mashable.com/2012/10/17/klout-perks-update/ (visitado 25-07-2014).

[64] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2014. URL: http://www.R-project.org.

[65] Everett M Rogers. *Diffusion of innovations*. Simon y Schuster, 2010.

[66] Kelly Ryan. *Twitter Study - August 2009*. Inf. téc. Texas, United States of America: Pear Analytics, ago. de 2009. URL: http://www.pearanalytics.com/wp-content/uploads/2012/12/Twitter-Study-August-2009.pdf (visitado 29-08-2014).

[67] Scott Sadowsky y Ricardo MartÃnez Gamboa. *Lista de Frecuencias de Palabras del Castellano de Chile (Lifcach)*. 2012. URL: http://sadowsky.cl/lifcach.html (visitado 29-08-2014).

[68] David Salomon. *Data compression: the complete reference*. Springer Science & Business Media, 2004, pág. 241. ISBN: 978-1-84628-602-5.

[69] Xiance Si y Maosong Sun. "Tag-LDA for scalable real-time tag recommendation". En: *Journal of Computational Information Systems* 6.1 (2009), págs. 23-31.

[70] Arlei Silva y col. "ProfileRank: Finding Relevant Content and Influential Users Based on Information Diffusion". En: *Proceedings of the 7th Workshop on Social Network Mining and Analysis*. SNAKDD '13. New York, NY, USA: ACM, 2013, 2:1-2:9. ISBN:

978-1-4503-2330-7. DOI: 10.1145/2501025.2501033. URL: http://doi.acm.org/10.1145/2501025.2501033.

[71] SUBTEL. *Informe Sectorial: Telecomunicaciones en Chile*. Inf. téc. Santiago, Chile: Ministerio de Transportes y Telecomunicaciones, mar. de 2013.

[72] SUBTEL. *Posicionamiento de Chile en Desarrollo Digital*. 2015. URL: http://www.subtel.gob.cl/attachments/article/5521/PPT%20Series%20Diciembre%202014%20VFinal.pdf (visitado 10-04-2015).

[73] Esther Velarde Consoli y Ricardo Canales G y Magali Meléndez J y Susana Lingán H. "Relación entre los procesos psicológicos de la escritura y el nivel socieconómico de estudiantes del Callado: Elaboración y baremación de una prueba de escritura de orientación cognitiva." En: *Investigación Educativa* 16.29 (2014). URL: http://revistasinvestigacion.unmsm.edu.pe/index.php/educa/article/view/7639.

[74] Zaiyong Tang, Chrys de Almeida y Paul A Fishwick. "Time series forecasting using neural networks vs. Box-Jenkins methodology". En: *Simulation* 57.5 (1991), págs. 303-310.

[75] Yee Whye Teh y col. "Hierarchical Dirichlet Processes". En: *Journal of the American Statistical Association* 101.476 (2006), págs. 1566-1581. DOI: 10.1198/016214506000000302. eprint: http://dx.doi.org/10.1198/016214506000000302. URL: http://dx.doi.org/10.1198/016214506000000302.

[76] Topsy. *Influence: the New Currency of the Web*. 2009. URL: http://web.archive.org/web/20090922223917/http://labs.topsy.com/influence (visitado 29-08-2014).

[77] Sandra Trafilaf. *Prueba Simce de escritura confirma incidencia de brecha socioeconómica, diarioUchile*. 2014. URL: http://radio.uchile.cl/2014/10/08/prueba-simce-de-escritura-confirma-incidencia-de-brecha-socioeconomica (visitado 28-01-2015).

[78] Twitter. *Getting started with Twitter*. 2014. URL: https://support.twitter.com/articles/215585# (visitado 29-08-2014).

[79] Twitter. *REST APIs*. URL: https://dev.twitter.com/rest/public (visitado 16-11-2014).

[80] Twitter. *The Streaming APIs*. URL: https://dev.twitter.com/streaming/overview (visitado 16-11-2014).

[81] Twitter. *Twitter Reports Second Quarter 2014 Results*. 2014. URL: https://investor.twitterinc.com/releasedetail.cfm?ReleaseID=862505 (visitado 29-10-2014).

[82] Twitter. *Using hashtags on Twitter*. URL: https://support.twitter.com/articles/49309-using-hashtags-on-twitter (visitado 16-11-2014).

[83] Yi Wang y col. "Plda: Parallel latent dirichlet allocation for large-scale applications". En: *Algorithmic Aspects in Information and Management*. Springer, 2009, págs. 301-314.

[84] Jianshu Weng y col. "TwitterRank: Finding Topic-sensitive Influential Twitterers". En: *Proceedings of the Third ACM International Conference on Web Search and Data Mining*. WSDM '10. New York, New York, USA: ACM, 2010, págs. 261-270. ISBN: 978-1-60558-889-6. DOI: 10.1145/1718487.1718520. URL: http://doi.acm.org/10.1145/1718487.1718520.

[85] Yusuke Yamamoto. *Twitter4j: A Java library for the Twitter API*. 2007. URL: http://www.twitter4j.org/.

[86] Min-Chul Yang y Hae-Chang Rim. "Identifying interesting Twitter contents using topical analysis". En: *Expert Systems with Applications* 41.9 (2014), págs. 4330 -4336. ISSN: 0957-4174. DOI: http://dx.doi.org/10.1016/j.eswa.2013.12.051. URL: http://www.sciencedirect.com/science/article/pii/S0957417414000141.

[87] Sarita Yardi, Daniel Romero, Grant Schoenebeck y col. "Detecting spam in a twitter network". En: *First Monday* 15.1 (2009).

[88] Ke Zhai y col. "Mr. LDA: A flexible large scale topic modeling package using variational inference in mapreduce". En: *Proceedings of the 21st international conference on World Wide Web*. ACM. 2012, págs. 879-888.

[89] WayneXin Zhao y col. "Comparing Twitter and Traditional Media Using Topic Models". English. En: *Advances in Information Retrieval*. Ed. por Paul Clough y col. Vol. 6611. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2011, págs. 338-349. ISBN: 978-3-642-20160-8. DOI: 10.1007/978-3-642-20161-5_34. URL: http://dx.doi.org/10.1007/978-3-642-20161-5_34.

Apéndice A

Resultados TS-LDA

En esta sección del apéndice se adjuntan los tópicos generados por TS-LDA con 100, 200 y 500 tópicos con sus 5 palabras más probables ordenados por el puntaje asociado de interés.

A.1. Cien tópicos

Tabla A.1: Aplicación de LDA con 100 tópicos

Puntaje	Tópico	Palabra 1	Palabra 2	Palabra 3	Palabra 4	Palabra 5
10.36	63	vin	gtgt	renac	arriend	mar
2.64	89	b	llam	san	esq	fueg
2.61	93	cas	pent	udi	error	polit
2.34	40	sur	accident	inform	sector	rut
2.29	99	reform	vot	diput	trabaj	polit
2.13	11	salud	favor	dm	envi	dat
2.11	13	incendi	bomber	forestal	comun	lug
1.92	38	region	santiag	in	metropolitan	valparais
1.7	23	feliz	navid	ano	cumplean	tod
1.6	72	com	ric	pan	cocin	almuerz
1.52	57	president	bachelet	punt	gobiern	entreg
1.5	46	te	amo	suen	extran	cumpl
1.4	37	escuch	cancion	music	ener	radi
1.23	64	col	camiset	don	cc	mon
1.19	5	estudi	result	univers	colegi	nacional
1.11	82	sal	mari	juan	sant	cruz
1.11	9	muert	atent	terror	ataqu	via
1.1	42	larrain	cas	justici	martin	mat
1.02	1	veran	program	inici	festival	centr
1.02	76	part	gol	jug	alexis	sanchez
0.97	51	ti	abraz	felic	graci	bes

Puntaje	Tópico	Palabra 1	Palabra 2	Palabra 3	Palabra 4	Palabra 5
0.85	74	part	u	fech	equip	libert
0.85	32	rob	carabiner	jov	avion	sigl
0.84	98	nivel	mayor	registr	caid	merc
0.82	31	tom	agu	cort	pel	beb
0.73	24	amor	llev	vid	corazon	viv
0.71	0	ley	medi	clas	proyect	present
0.69	81	fot	public	acab	facebook	nuev
0.68	68	sub	baj	viaj	pes	metr
0.67	50	problem	internet	servici	empres	funcion
0.62	39	salud	ministr	abort	dec	mentir
0.58	92	conoc	quier	viv	social	red
0.58	77	acuerd	eeuu	relacion	cub	chin
0.57	33	te	libr	grand	invit	leer
0.56	96	mejor	mund	peor	vid	ser
0.53	10	regal	compr	navid	naviden	perfect
0.52	49	chil	pais	argentin	chilen	visit
0.51	52	via	descarg	amp	ft	by
0.43	97	busc	perr	encontr	perd	gat
0.39	26	nuev	seguidor	unfollowers	termin	s
0.39	84	segund	prim	minut	dur	dak
0.37	75	amig	mis	tod	famili	junt
0.31	44	lt	tan	hermos	lind	encant
0.31	25	man	puert	dej	cerr	senor
0.3	22	necesit	vend	web	pagin	siti
0.16	35	hombr	muj	mujer	respet	parej
0.14	34	mal	calor	sol	fri	onda
0.14	66	graci	much	segu	x	te
0.12	55	manan	dorm	hor	qued	despert
0.12	4	lleg	cas	hor	esper	sal
0.07	85	te	esper	ver	estas	invit
0.05	62	hrs	manan	sab	lun	viern
0.04	65	dia	buen	teng	lind	seman
0.01	12	fot	gt	play	detall	revis
-0.1	94	gan	entrad	dan	p	premi
-0.14	43	pas	eso	sup	rap	piol
-0.16	69	hab	form	part	ide	cre
-0.24	14	termin	final	ver	capitul	tempor
-0.24	78	ver	pelicul	viend	tv	program
-0.29	86	cambi	vid	hay	person	nombr
-0.3	19	sient	odi	duel	cabez	dolor
-0.37	27	da	dio	igual	negr	mied
-0.38	95	tem	pobr	not	notici	e
-0.47	20	vuelv	vacacion	seman	vien	mes
-0.51	88	recuerd	tien	ven	razon	tod

Puntaje	Tópico	Palabra 1	Palabra 2	Palabra 3	Palabra 4	Palabra 5
-0.52	79	habl	pregunt	mand	llam	respond
-0.53	54	mam	viej	pap	pid	ped
-0.56	18	estar	deb	deberi	quier	contig
-0.6	16	son	tod	mejor	mis	esos
-0.65	87	anos	nin	hij	padr	muer
-0.65	45	mir	llor	sent	olvid	enamor
-0.78	3	gran	sea	tod	esper	grand
-0.81	17	gust	vide	list	agreg	reproduccion
-0.85	29	xd	jajajaj	jajajajaj	jajajajajaj	jajajajajajaj
-0.85	67	ano	nuev	pued	verl	vist
-0.88	41	car	pon	raj	pus	pat
-0.9	28	chic	pa	pal	cabr	tir
-0.91	61	cre	ser	val	dic	har
-0.96	71	the	of	i	on	to
-0.98	47	he	ultim	han	hech	h
-1.01	8	madr	real	club	jueg	mundial
-1.09	80	nuestr	tod	salud	bienven	cuent
-1.18	56	qued	peg	igual	falt	c
-1.2	59	sig	te	vuelt	c	twitt
-1.31	48	seri	eso	fuer	tan	igual
-1.32	90	pens	teni	dij	habi	eso
-1.41	60	tus	mejor	amig	twitt	experient
-1.43	2	jajaj	jaj	igual	xd	sup
-1.43	70	tod	moment	hay	estos	lad
-1.45	30	les	dej	dig	eso	falt
-1.54	58	cag	wn	sac	pa	mierd
-1.63	53	vez	cos	otra	primer	person
-2	7	weon	culia	viej	hueon	pur
-2.01	83	jajaj	xd	jajajaj	eso	po
-2.13	15	esa	wea	put	mierd	min
-2.29	91	hay	gent	entiend	tant	cre
-2.62	73	quier	te	twitt	descubr	ver
-2.96	21	buen	dias	noch	dia	tod
-4.27	6	q	d	x	xq	derech
-8.66	36	cuent	pag	mil	millon	plat

Fuente: Elaboración propia

A.2. Doscientos tópicos

Tabla A.2: Aplicación de LDA con 200 tópicos

Puntaje	Tópico	Palabra 1	Palabra 2	Palabra 3	Palabra 4	Palabra 5
22.98	93	person	excelent	vin	vist	info
14.72	78	vin	mar	diari	jardin	arriend
3.64	126	feliz	navid	cumplean	tod	famili
3.57	155	incendi	bomber	forestal	sector	lug
3.56	167	ministr	abort	molin	salud	renunci
3.44	123	pued	nuev	vist	gtgt	verl
3.42	163	b	esq	fueg	llam	basur
3.25	65	atent	ataqu	terror	charli	franci
3.23	57	cas	pent	carl	declar	velasc
3.13	50	larrain	martin	justici	atropell	juici
3.1	83	diput	reform	comision	senador	proyect
2.79	109	sur	sector	rut	nort	pist
2.25	165	ano	nuev	feliz	sea	celebracion
2	80	cas	udi	pent	silv	pid
1.83	128	sig	te	vuelt	estas	ver
1.77	47	plan	regional	alcald	intendent	realiz
1.64	1	region	in	santiag	metropolitan	at
1.61	107	baj	preci	pes	caid	cobr
1.52	26	com	ric	pan	chocolat	carn
1.51	22	campan	polit	financi	aport	fals
1.36	13	dak	etap	chilen	mot	segund
1.28	99	accident	lesion	dej	lug	vehicul
1.17	19	social	red	necesit	marc	gtgt
1.11	142	gust	vide	lik	official	oficial
1.03	24	ano	nuev	fiest	celebr	empez
0.99	154	gan	premi	mejor	concurs	oro
0.97	73	metr	estacion	central	fuert	leon
0.93	60	san	luis	pedr	jos	antoni
0.92	72	escuch	music	radi	prim	aguant
0.92	134	te	amo	extran	dir	matt
0.91	77	ley	medi	proyect	chil	tien
0.89	6	polit	pais	clas	chilen	interes
0.82	64	rob	sigl	carabiner	deten	delincuent
0.79	174	buen	noch	dia	descans	anim
0.78	198	te	invit	segu	grand	chil
0.76	160	felic	llen	dese	fiest	exit
0.76	90	president	bachelet	gobiern	piner	ex
0.75	3	jug	flor	part	cc	tapi
0.72	190	busc	curs	practic	desarroll	empres
0.7	104	estudi	univers	colegi	nacional	educacion

Puntaje	Tópico	Palabra 1	Palabra 2	Palabra 3	Palabra 4	Palabra 5
0.64	116	laboral	trabaj	reform	carg	gestion
0.61	175	cancion	escuch	ener	quier	conciert
0.57	55	ena	error	von	baer	val
0.57	12	sub	chil	bus	brasil	pasaj
0.56	136	pag	banc	plat	cobr	sueld
0.5	95	ener	sab	lun	manan	viern
0.44	112	necesit	maquillaj	tratamient	pein	limpiez
0.43	70	centr	cultur	santiag	parqu	aric
0.39	197	falt	respet	libert	expresion	limit
0.39	15	problem	internet	servici	solucion	tecnic
0.39	68	gran	grand	abraz	puebl	tremend
0.39	166	gol	alexis	sanchez	arsenal	city
0.38	168	u	part	union	gol	ohiggins
0.38	176	graci	segu	te	follow	devuelv
0.37	143	unid	eeuu	guerr	estad	cub
0.37	171	veran	play	disfrut	vacacion	activ
0.35	140	sal	uc	catol	univers	robert
0.35	42	cort	pel	luz	comun	energi
0.34	130	deb	estar	deberi	hab	hac
0.33	85	regal	navid	estas	perfect	naviden
0.33	150	consum	sufr	ejercici	efect	estres
0.31	199	perr	busc	encontr	perrit	gat
0.31	66	quier	viv	conoc	particip	ftisland
0.3	124	te	suen	sig	amo	mis
0.3	91	chil	pais	demand	cnn	bolivi
0.29	161	col	camiset	mon	part	iquiqu
0.21	81	dia	prim	manan	ser	empez
0.2	169	tom	beb	cervez	agu	caf
0.19	9	mejor	mund	peor	histori	eleg
0.16	92	amor	lt	vid	prueb	corazon
0.16	21	notici	period	not	prens	castr
0.16	4	amig	mis	companer	favorit	tod
0.16	69	present	festival	teatr	pastor	sot
0.13	94	capitul	final	teleseri	seri	turc
0.1	182	futbol	chilen	club	air	deport
0.1	54	seman	vien	proxim	mes	vacacion
0.1	97	fond	nuev	e	dispon	cen
0.09	84	pelicul	ver	cin	pobr	viend
0.07	32	dorm	manan	hor	despert	qued
0.07	20	roj	tia	jorg	daniel	rodrig
0.05	96	nacional	fech	primer	internacional	torne
0.02	18	calor	fri	agu	sec	piscin
0.02	193	segur	consej	entreg	evit	cuid
0.01	195	seri	fuer	gran	personaj	herman

Puntaje	Tópico	Palabra 1	Palabra 2	Palabra 3	Palabra 4	Palabra 5
0	33	libr	pap	francisc	leer	especial
-0.03	172	buen	dias	dia	teng	seman
-0.04	103	avion	cuerp	desaparec	negr	encontr
-0.05	29	lleg	cas	brav	esper	hug
-0.06	37	orden	rop	cas	ban	piez
-0.08	30	hombr	muj	mujer	tierr	sex
-0.1	185	car	raj	mentir	peor	pat
-0.14	43	puert	cerr	lag	montt	cierr
-0.17	40	plaz	juan	herrer	mall	johnny
-0.2	181	pas	rap	eso	piol	rat
-0.24	101	gan	dan	pon	viaj	dieron
-0.24	113	tem	nuev	amer	for	chil
-0.24	62	final	segund	prim	termin	tempor
-0.25	191	vid	mejor	import	histori	salv
-0.27	35	via	chin	pais	venezuel	situacion
-0.31	145	pal	senor	pic	ojo	vien
-0.32	87	compr	vend	entrad	k	convers
-0.32	48	web	public	siti	pagin	googl
-0.34	79	via	angel	mit	perez	plant
-0.34	120	sup	disc	show	band	rey
-0.35	147	qued	minut	hor	dur	falt
-0.37	45	gt	detall	fot	revis	imagen
-0.38	189	lind	hermos	bell	dia	lt
-0.4	17	hor	public	ultim	mencion	tweets
-0.4	59	viej	pascuer	viejit	port	agu
-0.41	158	les	dej	cont	parec	gust
-0.41	119	mis	mejor	son	seguidor	graci
-0.41	67	ciud	seren	visit	pase	via
-0.42	44	llev	viv	quer	mejor	cant
-0.43	111	tan	lind	bonit	ve	dificil
-0.45	188	mal	onda	buen	suert	tan
-0.47	138	nombr	alto	puent	leo	osorn
-0.5	152	madr	real	part	espan	viv
-0.51	8	sal	call	camin	salg	marcel
-0.52	141	llam	atencion	culp	dic	torr
-0.52	183	just	podr	ver	ser	fras
-0.52	153	hij	padr	nin	tendr	mes
-0.56	115	color	negr	blanc	mod	vest
-0.56	75	lleg	argentin	rodriguez	acuerd	juli
-0.58	23	cre	ser	har	unic	tendr
-0.58	7	fot	sac	jav	sub	instagram
-0.59	11	cambi	tien	encontr	ingres	entra
-0.59	74	via	out	mostrador	biobiochil	is
-0.61	63	volv	toc	pas	futur	dej

Puntaje	Tópico	Palabra 1	Palabra 2	Palabra 3	Palabra 4	Palabra 5
-0.61	117	ser	vidal	comentari	artur	cit
-0.64	149	part	form	unic	pid	perdon
-0.66	39	luc	ven	cas	vecin	carret
-0.66	186	acab	ver	corr	enter	escap
-0.66	156	ultim	han	h	tweet	retwitt
-0.67	192	tod	lad	dias	junt	igual
-0.68	98	pa	cag	qued	tod	peg
-0.7	118	sient	cabez	duel	dolor	maldit
-0.73	127	fot	facebook	nuev	public	he
-0.74	71	list	vide	i	reproduccion	he
-0.75	38	salud	tod	famili	aburr	especial
-0.76	36	via	amp	ft	by	descarg
-0.79	137	sol	bail	sal	mor	top
-0.83	56	mir	ti	ojos	hay	secret
-0.84	132	nivel	ment	test	personal	ideal
-0.86	196	nuev	otros	vent	servici	empres
-0.86	108	recuerd	olvid	cas	chef	matrimoni
-0.87	105	celul	iphon	carg	control	g
-0.87	52	tir	dej	rio	bio	peg
-0.87	25	habl	dej	trat	corazon	eso
-0.88	51	favor	mand	envi	x	mensaj
-0.91	179	ric	bes	boc	min	pobr
-0.98	121	derech	e	inform	human	charl
-1	194	punt	perd	estuv	aren	parec
-1.01	146	luch	logr	bienven	espaci	not
-1.04	27	min	paul	camil	cae	pele
-1.04	82	pas	mejor	vid	ide	pud
-1.05	131	hay	son	manan	santiag	condicion
-1.05	180	tod	nuestr	famili	amig	apoy
-1.06	28	canal	tv	program	lanz	youtub
-1.06	31	eso	ex	pabl	encant	jef
-1.08	135	com	hambr	uu	mayor	xd
-1.08	88	xd	jajaj	jajajaj	jajajajaj	jaj
-1.09	76	cuent	dar	di	twitt	cre
-1.11	157	pens	habi	teni	sabi	iba
-1.14	159	chic	cabr	pon	herman	wen
-1.15	125	man	ayud	met	selfi	sac
-1.15	144	vez	otra	primer	chil	veo
-1.15	53	da	pen	mied	dio	paj
-1.21	10	anos	muer	jov	mat	busc
-1.24	148	mil	millon	pes	vec	dolar
-1.24	14	son	estos	mism	esos	estas
-1.27	58	habl	pur	escrib	weas	gent
-1.27	177	he	sid	hech	hub	han

Puntaje	Tópico	Palabra 1	Palabra 2	Palabra 3	Palabra 4	Palabra 5
-1.3	178	tus	mejor	amig	twitt	experient
-1.36	34	quier	sant	mari	cruz	ros
-1.36	114	the	gust	vide	of	trail
-1.38	106	u	part	azul	jug	pat
-1.41	173	quier	ver	estar	estadi	viend
-1.43	129	llor	hac	sent	trist	par
-1.48	49	igual	sup	xd	razon	jaj
-1.53	187	anos	nin	cumpl	celebr	edad
-1.56	170	c	alcanz	manan	maxim	vient
-1.59	133	esper	sea	mejor	ano	dia
-1.6	100	encuentr	vuelv	loc	ciudadan	via
-1.66	184	person	esa	hay	dec	cre
-1.76	2	te	quier	ti	contig	conmig
-1.86	151	mam	dij	herman	dic	queri
-1.88	41	moment	estos	dias	cualqui	oportun
-1.91	164	don	gan	xd	pong	vol
-2.06	86	termin	m	s	viaj	espanol
-2.06	110	culia	mierd	culi	weon	hueon
-2.07	102	hrs	tod	pierd	tuit	quier
-2.16	16	quier	te	twitt	descubr	alta
-2.16	89	gent	hay	odi	entiend	carg
-2.41	0	cos	hay	vec	much	tant
-2.58	46	wea	put	esa	mierd	weon
-2.7	162	pregunt	dic	eso	dig	import
-2.81	122	justin	one	sig	harry	re
-3.82	139	wn	po	ctm	weon	xd
-4.57	5	q	hay	dic	cre	dec
-5.1	61	d	q	x	n	cn

Fuente: Elaboración propia

A.3. Quinientos tópicos

Tabla A.3: Aplicación de LDA con 500 tópicos

Puntaje	Tópico	Palabra 1	Palabra 2	Palabra 3	Palabra 4	Palabra 5
18.79	364	vin	mar	diari	jardin	gtgt
18.35	379	person	excelent	vist	vin	info
9.05	43	b	esq	llam	fueg	basur
8.72	274	incendi	bomber	forestal	sector	alert
7.56	465	dm	dat	salud	contact	favor
7.52	310	sector	sur	rut	vehicul	nort
6.47	280	pent	cas	declar	delan	fiscali
6.45	49	region	in	santiag	metropolitan	at
5.32	373	nuev	seguidor	unfollowers	estadistE	yorkE
5.2	107	terror	charli	atent	franci	ataqu
5.17	416	pued	nuev	verl	vist	gtgt
5.14	139	ministr	abort	molin	renunci	heli
4.83	213	larrain	martin	justici	juici	conden
4.78	387	cas	udi	pent	silv	polit
4.72	193	social	red	necesit	gtgt	contactan
4.22	144	estudi	univers	carrer	psu	nacional
4.1	101	u	gol	gutierrez	uc	wanderers
3.96	13	san	pedr	luis	antoni	marc
3.84	111	te	estas	ver	invit	esper
3.8	335	proyect	regional	intendent	entreg	ministr
3.77	330	bachelet	campan	president	financi	aport
3.56	246	via	descarg	ft	amp	by
3.55	216	reform	laboral	gobiern	critic	agend
3.55	363	dak	etap	mot	chilen	general
3.37	461	feliz	navid	ano	dese	tod
3.23	409	conoc	viv	quier	particip	ftisland
3.2	148	vot	sistem	binominal	senador	parlamentari
3.13	171	error	ena	von	baer	moreir
3	118	fot	public	acab	he	facebook
2.99	235	buen	dia	teng	dias	excelent
2.95	2	preci	cobr	baj	caid	merc
2.92	126	ano	nuev	feliz	abraz	celebracion
2.84	186	ano	nuev	feliz	sea	exit
2.83	249	col	mon	camiset	coc	iquiqu
2.7	228	com	poll	carn	ric	cocin
2.69	196	noch	buen	descans	dorm	duerm
2.6	85	comision	educacion	reform	proyect	educacional
2.59	48	baj	sub	pes	preci	pasaj
2.58	137	polit	pais	chilen	clas	chil
2.56	112	rob	sigl	delincuent	deten	carabiner

Puntaje	Tópico	Palabra 1	Palabra 2	Palabra 3	Palabra 4	Palabra 5
2.44	344	ano	nuev	celebr	fiest	comenz
2.31	177	gust	vide	official	hd	espanol
2.3	333	ultim	h	tweets	tweet	han
2.19	362	km	sism	corr	intens	profund
2.16	259	necesit	perrit	ayud	favor	busc
2.11	14	ley	proyect	sot	pastor	present
2.03	157	necesit	maquillaj	tratamient	pein	limpiez
2.03	242	agu	tom	cervez	hel	beb
2.02	457	futbol	equip	jug	chilen	jugador
2.02	83	medi	ley	clas	digital	maraton
1.95	285	hrs	viern	ener	sab	lun
1.91	380	chil	sub	rodriguez	brasil	argentin
1.9	389	escuch	radi	music	fm	program
1.87	257	com	chocolat	hel	tort	ric
1.82	19	sig	vuelt	te	twitt	dak
1.64	265	duel	cabez	dolor	espald	guat
1.61	86	calor	fri	cap	lluvi	cag
1.6	161	juan	herrer	pabl	johnny	vecchi
1.59	271	curs	veran	escuel	activ	tall
1.56	97	millon	mil	pag	pes	dolar
1.54	470	madr	hij	put	padr	real
1.52	492	bus	viaj	estacion	tren	terminal
1.51	21	festival	teatr	artist	confirm	present
1.47	291	suen	te	mis	amo	sig
1.45	322	cort	pel	luz	energi	electr
1.44	300	flor	jug	tapi	defens	felip
1.42	353	araucani	temuc	gobiern	mapuch	carabiner
1.38	397	list	vide	he	reproduccion	agreg
1.33	467	amor	paz	vid	etern	prueb
1.33	155	metr	oro	leon	cristian	balon
1.32	452	feliz	navid	cumplean	regal	celebr
1.31	159	tom	caf	decision	desayun	ram
1.3	337	inform	tecnic	servici	personal	nuestr
1.29	395	gonzalez	jorg	rios	uc	rodrig
1.27	441	llam	atencion	numer	telefon	contest
1.26	207	son	mejor	mis	quien	e
1.26	108	fot	nuev	facebook	publiqu	perfilE
1.24	166	ener	edicion	sab	n^{o}	turismocuatr
1.24	239	madur	venezuel	cub	eeuu	pres
1.24	9	iphon	googl	andro	apple	aplic
1.23	50	puert	cerr	montt	var	lag
1.23	138	libr	air	leer	acondicion	leyend
1.21	390	segur	entreg	consej	recomend	campan
1.2	269	dorm	despert	hor	suen	despiert

Puntaje	Tópico	Palabra 1	Palabra 2	Palabra 3	Palabra 4	Palabra 5
1.18	174	maxim	minim	°c	manan	grad
1.17	474	avion	desaparec	busqued	recuper	airasi
1.17	176	libert	expresion	fanat	cop	limit
1.16	172	dej	muert	her	accident	person
1.15	398	dia	lind	buen	teng	noch
1.15	203	hombr	muj	mujer	sex	sexual
1.14	133	com	ric	pastel	almuerz	chocl
1.14	66	gan	concurs	entrad	dobl	particip
1.14	115	nin	anos	registr	jo	onur
1.1	400	seri	personaj	descubr	ideal	fuer
1.08	247	fot	jav	aceved	filtr	instagram
1.08	80	primer	fech	vez	segund	nacional
1.07	345	i	to	the	lov	you
1.06	178	escuch	cancion	quier	conciert	ener
1.06	180	paul	mast	chef	cocin	abuelit
1.04	499	recuerd	event	llen	tien	matrimoni
1.04	160	buen	dias	tard	noch	tl
1.02	418	internet	servici	senal	conect	funcion
1.01	65	tv	canal	program	tvn	notici
1	110	gan	premi	manuel	increibl	city
0.99	404	seman	proxim	febrer	marz	vacacion
0.97	343	buen	noch	dias	dia	madrug
0.97	410	car	raj	pat	pot	cul
0.95	153	pelicul	ver	cin	viend	estren
0.94	444	mejor	pelicul	oscar	nomin	actor
0.92	81	gt	detall	revis	nuev	marc
0.92	460	amer	espanol	latin	cop	for
0.89	75	chil	pais	nivel	desarroll	capital
0.88	382	cultur	centr	arte	cultural	muestr
0.84	188	mat	jov	hombr	asesin	muert
0.84	357	viej	viejit	pascuer	regal	port
0.84	340	te	amo	dir	matt	lt
0.83	305	negr	blanc	camiset	poler	color
0.81	374	unid	hospital	medic	estad	curic
0.81	106	link	aguant	prim	dari	singl
0.8	447	final	capitul	tempor	ver	vi
0.76	296	disc	band	present	music	rock
0.75	438	amig	mejor	twitt	tus	son
0.74	243	gust	vide	trail	vardoc	grab
0.72	406	dia	buen	alegr	inocent	enoj
0.71	277	derech	mit	human	izquierd	perez
0.7	73	mund	mejor	mundial	bbc	enter
0.7	198	amig	mis	companer	mejor	polol
0.7	328	gran	felicit	chilen	tremend	exit

Puntaje	Tópico	Palabra 1	Palabra 2	Palabra 3	Palabra 4	Palabra 5
0.69	11	cambi	horari	vid	ingres	sal
0.69	448	graci	much	felic	ti	abraz
0.67	327	perr	gat	human	mascot	cuid
0.64	478	aut	manej	mat	choc	conduc
0.62	217	ban	rop	orden	piez	limpi
0.61	302	u	chil	univers	estadi	hinch
0.61	396	ti	igual	graci	tb	confi
0.6	202	trabaj	empres	contrat	realiz	import
0.6	286	estar	estadi	foo	haz	nacional
0.6	375	mal	suert	cue	practic	tan
0.58	102	sal	mari	jos	sant	entrev
0.52	324	dan	gan	asco	color	veo
0.51	55	alto	parqu	puent	pase	vuel
0.5	393	deb	estar	imagin	hac	admit
0.49	185	veran	program	disfrut	vacacion	entreten
0.48	123	beb	alcohol	consum	efect	estres
0.48	63	onda	mal	buen	ide	gent
0.48	472	corazon	romp	guard	silenci	llen
0.47	93	ex	president	piner	sebasti	novi
0.46	95	habi	olvid	pens	dad	dich
0.46	279	hab	deb	deberi	debi	pud
0.46	350	jueg	jug	ps	lol	pelot
0.44	54	pap	francisc	critic	mis	hij
0.43	240	bail	cant	cancion	escuch	cumbi
0.42	431	pid	ped	perdon	disculp	olla
0.42	179	municipal	column	comun	alcald	provident
0.4	391	pic	ojo	boc	met	tap
0.4	414	vid	larg	salv	mia	histori
0.39	498	fot	sub	set	instagram	subi
0.38	91	plaz	mall	seren	armas	itali
0.37	82	hor	juli	ultim	medi	ult
0.35	476	hub	sid	habri	hech	estari
0.34	309	les	gust	dej	cuest	molest
0.34	167	fueg	artificial	torr	show	espectacul
0.33	223	estar	deberi	deb	atent	durm
0.33	436	tan	dificil	facil	simpl	seri
0.32	386	dud	consult	ayud	salud	futur
0.31	169	da	mied	paj	verguenz	ris
0.3	41	nacional	chil	bar	er	congres
0.3	20	viv	llev	quer	estudi	amor
0.3	158	chil	puebl	final	tenis	avanz
0.29	190	teleseri	turc	chilen	pucon	novel
0.29	131	punt	aren	sum	perd	estuv
0.27	316	tem	interes	opinion	column	opin

Puntaje	Tópico	Palabra 1	Palabra 2	Palabra 3	Palabra 4	Palabra 5
0.27	100	via	escuch	amp	salt	tem
0.27	402	chil	salud	cnn	exist	vendr
0.26	94	pan	com	pascu	pet	isla
0.25	78	tierr	amarill	ciel	estrell	planet
0.25	51	play	piscin	sol	veran	disfrut
0.23	326	pag	verd	impuest	compr	tarjet
0.23	462	famili	disfrut	amig	compart	tod
0.22	17	mejor	pierd	cortin	roll	unidad
0.22	321	dia	prim	ano	termin	comenz
0.2	28	vidal	matrimoni	artur	inter	medel
0.18	267	part	form	graci	unic	ano
0.16	325	termin	m	s	viaj	carrer
0.16	87	grand	exit	sos	maestr	idol
0.15	205	lun	manan	seman	viern	empiez
0.14	127	quer	ser	amig	tod	disfrut
0.11	220	tir	libr	piedr	mierd	chanch
0.11	183	andres	velasc	pag	almuerz	via
0.11	36	estas	perfect	regal	cam	profesional
0.11	411	pen	da	val	dio	vali
0.1	288	nort	eeuu	core	duen	via
0.09	294	falt	respet	poquit	merec	opinion
0.09	173	les	tod	dese	usted	pid
0.08	262	cag	ris	pa	pur	dej
0.06	4	ano	termin	empez	click	resum
0.04	264	te	segu	invit	chil	feri
0.02	319	desafi	calet	ray	volc	conquist
0.02	56	internacional	valdivi	triatlon	torne	ii
0.02	136	quier	twitt	te	descubr	alta
0.01	241	segund	part	comenz	comienz	tempor
-0.01	64	salud	tod	famili	estim	amig
-0.01	276	val	mia	ide	callamp	aric
-0.02	439	argentin	estudi	mexic	mes	program
-0.02	303	ser	dia	manan	tendr	habr
-0.02	314	dio	hambr	suen	muer	ris
-0.03	165	the	of	via	tempor	gam
-0.03	255	roj	fum	pep	dej	dieg
-0.04	37	sol	sal	piel	tom	quem
-0.05	437	busc	acab	encontr	utiliz	encuentr
-0.06	197	santiag	centr	viaj	stgo	aeropuert
-0.07	405	c	vient	alcanz	maxim	minim
-0.07	250	qued	dorm	hor	ire	queri
-0.07	428	call	sal	camin	grit	cent
-0.08	287	leer	comentari	te	ser	invit
-0.08	251	via	guerr	palestin	mundial	israel

Puntaje	Tópico	Palabra 1	Palabra 2	Palabra 3	Palabra 4	Palabra 5
-0.08	92	tan	tont	gent	estup	idiot
-0.09	227	inform	derech	e	guzm	das
-0.09	237	pal	hij	pic	temaz	vien
-0.09	323	cuerp	quit	prest	quem	pis
-0.11	238	habl	dej	contig	tem	ignor
-0.11	484	pens	iba	sabi	habi	estab
-0.11	234	luc	sal	top	via	barri
-0.11	236	toc	tem	guitarr	clasic	fond
-0.12	261	minut	hor	qued	poc	demor
-0.13	40	he	vist	dich	ten	vec
-0.14	469	logr	met	super	objet	impos
-0.14	150	otra	vez	chil	tour	wants
-0.15	289	favor	x	siguem	veo	ayud
-0.16	27	encontr	ingres	pagin	form	aument
-0.16	298	pap	especial	gran	felicit	imag
-0.17	125	mir	version	car	espej	imagen
-0.18	129	tod	lad	dej	oscur	esos
-0.19	256	pas	rap	tan	vol	crec
-0.19	451	esper	result	respuest	estes	ansi
-0.19	368	dia	manan	doming	descans	sab
-0.2	151	acuerd	firm	principi	chil	lleg
-0.2	372	sac	fot	selfi	sorprend	sonris
-0.2	453	respet	estudi	public	denunci	veran
-0.21	32	mayor	banc	cuent	x	adult
-0.21	371	cam	levant	acost	manan	paj
-0.21	1	pa	cag	cach	sirv	ven
-0.23	422	sac	chuch	crest	saqu	weon
-0.24	22	sant	quier	cruz	doming	laur
-0.24	29	te	canal	youtub	lanz	gust
-0.24	201	fiest	felic	estas	celebr	sean
-0.25	187	pat	rubi	part	patrici	marin
-0.26	299	odi	maldit	calor	mierd	resfri
-0.26	145	histori	cont	fom	chist	final
-0.27	290	perd	perdi	mied	oportun	tuv
-0.27	426	te	amo	lleg	gust	list
-0.28	283	disen	inclu	lent	pes	grafic
-0.28	486	brav	claudi	barcelon	messi	luis
-0.28	497	moment	cualqui	estos	oportun	lleg
-0.29	487	vez	primer	veo	otra	vi
-0.3	212	tod	fuerz	fe	nuestr	anim
-0.31	215	cos	cualqui	otra	esas	otras
-0.31	388	pas	piol	igual	sup	vol
-0.32	214	man	levant	met	propi	iron
-0.32	34	vien	prepar	manan	seman	tremend

Puntaje	Tópico	Palabra 1	Palabra 2	Palabra 3	Palabra 4	Palabra 5
-0.33	442	ver	acab	termin	enter	grab
-0.33	483	plan	etc	emergent	colegi	edifici
-0.34	96	dej	tranquil	entrar	respir	dejam
-0.34	12	min	mostr	polol	pot	calient
-0.34	408	period	chil	chilen	line	model
-0.35	348	peor	son	fals	dud	mentir
-0.35	7	mil	person	pes	gran	tablet
-0.35	446	lleg	cas	oficin	casit	har
-0.35	15	nuestr	tiend	local	descuent	compr
-0.36	156	culia	feo	weon	flait	maricon
-0.36	253	lleg	hor	esper	atras	demor
-0.37	3	not	falt	enferm	mental	fisic
-0.37	424	parej	relacion	amor	gay	leo
-0.39	306	ayud	graci	te	nuestr	tod
-0.39	113	senor	jesus	nac	nin	naci
-0.4	475	problem	solucion	hay	tuv	posibl
-0.4	399	son	cual	esas	mayori	esos
-0.41	420	qued	poquit	poc	fuer	cup
-0.41	346	notici	via	lee	resum	present
-0.41	459	mejor	eleg	opcion	mund	compani
-0.41	42	mis	favorit	amig	companer	vacacion
-0.42	421	teni	ide	sabi	razon	fe
-0.42	124	loc	vuelv	volvi	volv	nen
-0.43	119	llev	tendr	anos	hij	cuant
-0.44	383	nombr	cambi	llam	pon	apell
-0.44	69	pas	cos	rar	suel	suced
-0.44	114	just	vill	necesari	orig	caig
-0.45	90	import	vid	cos	eso	rest
-0.45	77	manan	am	hor	levant	ire
-0.45	315	peg	qued	jef	hac	comb
-0.46	5	bell	ciud	hermos	dia	lug
-0.46	204	mand	envi	mensaj	dm	corre
-0.47	132	sent	comun	humor	vid	hac
-0.47	31	dec	quis	deb	eso	queri
-0.47	260	part	termin	final	comenz	jueg
-0.47	270	maner	tod	mejor	empez	termin
-0.47	318	lleg	veng	cas	esper	avis
-0.47	229	palabr	sabi	clav	defin	dic
-0.48	493	vall	destin	new	tim	happy
-0.49	52	gener	pais	expert	chil	maestr
-0.49	392	vend	compr	entrad	vent	convers
-0.5	417	dej	llev	papel	bols	bot
-0.51	490	camin	sal	tac	cas	lleg
-0.51	266	marc	mejor	conoc	histori	leandr

Puntaje	Tópico	Palabra 1	Palabra 2	Palabra 3	Palabra 4	Palabra 5
-0.51	341	mentir	grand	mati	fernandez	nuestr
-0.52	419	trat	entend	empat	eso	ignor
-0.52	304	dar	andar	vuelt	dand	bici
-0.53	443	pobr	angelit	weon	tip	fach
-0.53	407	recuerd	olvid	record	tra	vid
-0.55	76	alexis	sanchez	gol	arsenal	varg
-0.56	301	te	extran	beach	socc	acuerd
-0.57	440	salud	x	lt	estim	pls
-0.58	263	junt	tod	vecin	amig	carret
-0.58	369	apoy	graci	salud	agradec	amig
-0.59	377	exist	cre	deberi	hay	gent
-0.59	477	hac	sig	seguir	tramit	reir
-0.59	168	ric	ve	min	mijit	cost
-0.6	59	mejor	sec	segur	rap	presupuest
-0.61	122	fot	corr	caj	escap	andes
-0.62	57	graci	much	dispon	ok	diari
-0.62	0	gan	dieron	fav	quit	ta
-0.63	334	muert	castr	fidel	muri	rumor
-0.63	339	han	hech	has	sid	ten
-0.64	10	tien	razon	tod	castig	encuentr
-0.64	53	amig	secret	hay	ti	regal
-0.65	199	ven	mejor	disfrut	vei	verm
-0.65	273	pas	can	horribl	peor	cel
-0.65	142	cas	lleg	abuel	qued	camp
-0.66	385	web	public	nuestr	siti	pagin
-0.67	219	luch	vid	jar	ejempl	victor
-0.68	141	sup	junior	ando	oye	fom
-0.69	366	gol	pared	suaz	fierr	penal
-0.71	349	tan	lind	tiern	encant	awww
-0.71	117	te	amo	jur	hermos	imagin
-0.73	403	mejor	vid	tod	calid	merec
-0.73	149	pur	weas	habl	qe	esas
-0.73	221	unic	cre	form	ser	pens
-0.74	99	tus	experient	usand	cuentan	erp
-0.75	143	control	situacion	complic	actual	trat
-0.75	208	via	fot	brindis	tumblr	ment
-0.75	195	via	vide	check	official	preview
-0.76	401	dur	carg	celul	mes	bateri
-0.77	181	anos	muer	edad	joe	muri
-0.77	342	escrib	parec	ejempl	histori	suen
-0.79	164	vuelv	encuentr	ciudadan	ener	via
-0.8	463	cas	sal	carret	lleg	sirv
-0.81	128	bio	plant	marihuan	via	cultiv
-0.81	412	person	tip	esas	conozc	hay

Puntaje	Tópico	Palabra 1	Palabra 2	Palabra 3	Palabra 4	Palabra 5
-0.82	293	wea	put	weon	wn	fom
-0.83	232	habl	aprend	ingles	chin	ensen
-0.83	370	anos	mes	cumpl	llev	tres
-0.83	252	via	fall	out	cienci	is
-0.85	295	hueon	hue	put	esa	pur
-0.85	320	graci	pued	da	encuest	colabor
-0.85	365	don	genar	hech	bad	doy
-0.87	473	vin	dia	melon	mejor	cat
-0.87	376	camil	pele	reality	prueb	eugeni
-0.87	8	viv	empez	ve	direct	sig
-0.88	194	mierd	weon	mand	mism	calor
-0.88	367	par	esper	mes	hor	reir
-0.9	162	tod	dias	andan	vay	caen
-0.9	445	funcion	normal	quint	horari	traduccion
-0.91	44	aburr	cans	toy	peg	chat
-0.92	423	cur	ke	john	diput	nacional
-0.93	23	bonit	tan	ve	car	lind
-0.93	464	anim	puch	mejor	lat	dibuj
-0.94	284	grup	mand	habl	whatsapp	fb
-0.95	146	vi	sali	pens	acord	teni
-0.96	496	fot	pon	pued	porf	d
-0.96	292	cre	deb	eso	equivoc	unic
-0.97	471	trist	realid	sient	vid	tan
-0.97	272	complet	falt	maravill	total	desastr
-0.98	105	parec	perr	tej	calz	per
-0.99	233	tod	usted	pa	sobr	bast
-0.99	488	pequen	grand	mund	gigant	diferent
-0.99	130	pas	rat	dej	too	vei
-1.01	434	tod	son	igual	sean	lady
-1.01	71	hay	son	condicion	santiag	dec
-1.02	307	color	pint	ros	pon	azul
-1.03	317	jajaj	xd	cach	jajajaj	igual
-1.05	154	plat	compr	diner	gast	pag
-1.05	38	jajajaj	xd	igual	jajajajajaj	chistos
-1.06	191	igual	bac	jajaj	jaj	bkn
-1.07	35	cas	lleg	arregl	pel	calm
-1.07	430	conoc	histori	cons	casual	detall
-1.08	134	wen	pa	terribl	po	ta
-1.08	25	tel	luz	apag	prend	viend
-1.08	62	chil	demand	moral	bolivi	acept
-1.08	89	dal	sea	import	llen	esper
-1.08	354	gent	entiend	carg	habl	esa
-1.09	192	via	present	bellez	increibl	taylor
-1.1	200	rey	via	antigu	leo	lan

Puntaje	Tópico	Palabra 1	Palabra 2	Palabra 3	Palabra 4	Palabra 5
-1.1	449	te	avis	preocup	vay	apuest
-1.1	331	llor	hiz	reir	emocion	hic
-1.13	489	tod	aparec	tl	veo	tuiter
-1.13	355	viej	culi	marac	fea	mierd
-1.14	210	te	fuist	extran	amar	ama
-1.14	313	angel	via	cc	montan	ricard
-1.14	258	te	vas	ves	dig	tien
-1.16	425	turn	salud	revis	doctor	pol
-1.17	6	seri	mejor	pud	tuv	evit
-1.18	103	pregunt	respond	respuest	hic	acept
-1.19	352	part	cer	ningun	aport	hay
-1.19	135	lt	respond	cit	tendri	estas
-1.19	480	son	mism	diferent	cos	distint
-1.19	147	cre	eso	hay	suficient	peor
-1.2	282	jajaj	igual	gtlt	eso	xd
-1.2	18	dej	habl	caer	pesc	quier
-1.2	104	quier	volv	ver	futur	realid
-1.21	72	podr	ser	c	ver	dorm
-1.21	61	sal	sali	salg	vacacion	trot
-1.22	359	lt	hermos	encant	lind	amor
-1.22	244	mejor	sea	esper	desped	sean
-1.23	378	sig	ofert	siguen	busc	dal
-1.24	429	culp	sufr	perd	tuv	asum
-1.24	381	mor	prefier	quier	see	qued
-1.25	170	k	bienven	dam	nuestr	cuent
-1.25	24	jajaj	obvi	po	jajajaj	xd
-1.26	281	mod	on	lin	estil	liv
-1.26	46	x	inici	espaci	ocup	tremend
-1.28	455	one	harry	niall	zayn	re
-1.29	58	jajajaj	ok	jajaj	xd	sorry
-1.31	468	pas	eso	cos	peor	veg
-1.38	413	jajaj	xd	igual	jajajaj	notabl
-1.4	312	seri	fuer	gran	herman	porfavor
-1.43	225	sea	esper	mejor	ano	gran
-1.43	308	tuit	anterior	popul	borr	copi
-1.44	224	fot	justin	bieb	model	mon
-1.44	120	oye	pas	dej	dient	yap
-1.48	479	parec	hiz	envidi	pur	famos
-1.49	79	vec	mil	visit	perfil	vari
-1.52	427	sient	tan	orgull	sent	rar
-1.52	311	pas	eso	parec	habr	cre
-1.53	351	jajaj	jajajaj	brom	noo	jaj
-1.57	456	conmig	enoj	quier	dim	ven
-1.59	67	pon	pus	pong	pil	nervios

Puntaje	Tópico	Palabra 1	Palabra 2	Palabra 3	Palabra 4	Palabra 5
-1.61	116	gord	veo	igual	flac	estas
-1.65	495	te	pierd	esper	acompan	oportun
-1.66	450	pon	gustari	rio	x	graci
-1.67	98	dij	queri	mam	iba	teni
-1.67	30	ver	quier	entren	ire	ed
-1.68	152	ud	salud	coment	opin	parec
-1.7	140	chic	cabr	herman	letr	grand
-1.71	45	fuert	temblor	eso	ruid	senti
-1.71	435	hay	gent	poc	much	molest
-1.73	39	p	jajaj	jajajaj	jajajajaj	jaj
-1.73	206	abraz	bes	ti	lind	gran
-1.74	415	deport	club	oficial	fans	lot
-1.76	16	hay	esper	cuid	ten	pacienci
-1.78	226	tant	hay	cos	gent	entiend
-1.78	68	pas	eso	cre	igual	peor
-1.81	358	dic	tia	dij	tio	yoli
-1.87	433	uu	c	queri	puch	igual
-1.87	218	jajaj	xd	parec	bloqu	jajajaj
-1.88	332	jaj	jajaj	igual	sii	eso
-1.94	70	estuv	viend	buenisim	gustav	rat
-1.95	482	jajaj	xd	jajajaj	igual	luli
-1.96	360	quier	contig	estar	volv	vert
-1.99	182	piens	igual	pens	eso	cre
-2.02	361	quier	ver	giorgianolmlE	entrarE	sepE
-2.07	268	eso	signif	pas	record	dic
-2.07	248	jajajaj	xd	jajajajaj	xdd	xddd
-2.13	209	anda	v	parec	suelt	oye
-2.13	175	hermos	lind	guap	precios	encant
-2.17	347	hay	gent	hart	darl	cos
-2.2	163	jajajajaj	xd	jajajajajaj	jajajajajajaj	siiii
-2.25	394	te	enamor	gust	ti	doy
-2.27	245	ctm	wn	cag	zorr	mierd
-2.27	184	dig	eso	dic	les	mient
-2.41	384	weon	po	oye	pa	estay
-2.42	109	esa	wea	mism	mierd	sensacion
-2.44	466	jaj	igual	po	eso	cach
-2.46	481	prens	fuent	sal	dic	conferent
-2.51	458	xd	jajajajajaj	csm	jajajajajajaj	jajajajaj
-2.6	297	xd	jajajaj	jajajj	jajajajj	jajajjaj
-2.65	485	quier	ver	irme	dec	uu
-2.65	336	hag	quier	sea	esper	tt
-2.72	26	xd	jajajaj	jajajajaj	hahah	hah
-2.75	211	cre	har	ser	herman	libr
-2.77	189	xd	ajajaj	igual	ajaj	ajajajaj

Puntaje	Tópico	Palabra 1	Palabra 2	Palabra 3	Palabra 4	Palabra 5
-2.83	60	jajaj	imagin	eso	encant	jaj
-2.86	278	q	d	pq	porq	dic
-2.86	84	d	carg	gestion	riesg	s
-2.87	74	q	cre	hay	dic	supon
-2.88	329	wn	po	xd	cag	pa
-2.89	47	esa	fras	cancion	actitud	encant
-2.95	230	dic	eso	entiend	pq	hay
-2.98	254	eso	entiend	refier	dic	dig
-3.07	356	cuent	di	dar	twitt	regres
-3.21	121	e	c	l	n	p
-3.31	275	graci	segu	te	follow	devuelv
-3.39	491	ojos	mir	bes	tus	mis
-3.58	88	q	x	d	xq	ej
-4.42	222	mam	herman	dic	hij	dij
-4.71	33	estos	son	dias	moment	wns
-5.54	494	son	esos	tod	mism	tip
-7.92	231	regal	navid	naviden	compr	arbol
-8.55	432	d	q	x	cn	n
-8.67	454	hay	otros	pais	diferent	cre
-9.61	338	hic	pas	eso	posibl	georg

Fuente: Elaboración propia

Apéndice B

Lista de Stopwords en Español

me	un	una	unas
unos	uno	sobre	todo
tambien	tras	otro	algun
alguno	alguna	algunos	algunas
ser	es	soy	eres
somos	sois	estoy	esta
estamos	estais	estan	como
en	para	atras	porque
por que	estado	estaba	ante
antes	siendo	ambos	pero
por	poder	puede	puedo
podemos	podeis	pueden	fui
fue	fuimos	fueron	hacer
hago	hace	hacemos	haceis
hacen	cada	fin	incluso
primero	desde	conseguir	consigo
consigue	consigues	conseguimos	consiguen
ir	voy	va	vamos
vais	van	vaya	gueno
ha	tener	tengo	tiene
tenemos	teneis	tienen	el
ella	la	lo	le
las	los	su	aqui
mio	tuyo	ellos	ellas
nos	nosotros	vosotros	vosotras
si	dentro	solo	solamente
saber	sabes	sabe	sabemos
sabeis	saben	ultimo	largo
bastante	haces	muchos	aquellos
aquellas	sus	entonces	tiempo
verdad	verdadero	verdadera	cierto

ciertos	cierta	ciertas	intentar
intento	intenta	intentas	intentamos
intentais	intentan	dos	bajo
arriba	encima	usar	uso
usas	usa	usamos	usais
usan	emplear	empleo	empleas
emplean	ampleamos	empleais	valor
muy	era	eras	eramos
eran	modo	bien	cual
cuando	donde	mientras	quien
con	entre	sin	trabajo
trabajar	trabajas	trabaja	trabajamos
trabajais	trabajan	podria	podrias
podriamos	podrian	podriais	yo
aquel	el	de	que
en	y	a	uno
del	se	por	con
su	no	para	al
este	como	el	mas
o	yo	otro	pero
todo	si	ese	entre
sin	ya	mucho	sobre
tambien	quien	desde	cuando
porque	tanto	hasta	solo
este	donde	mismo	nuestro
uno	mi	ademas	asi
cual	ese	todo	hoy
segun	durante	bien	ayer
cada	alguno	contra	ahora
que	tu	ni	despues
menos	luego	mucho	aunque
mientras	siempre	alguno	poco
ante	sino	tras	antes
nada	tal	aun	frente
algo	dentro	varios	bajo
hacia	si	como	cualquiera
pues	incluso	cuyo	aqui
aquel	nunca	casi	mas
entonces	cuanto	cerca	nadie
tarde	tu	ahi	ambos
ninguno	especialmente	claro	bueno
bastante	ninguno	asimismo	mediante
todavia	tampoco	ambos	finalmente
vamos	distinto	demas	actualmente

cual	quien	tanto	alguien
mismo	adelante	mucho	tal
alli	recien	diverso	dios
alla	nuevamente	principalmente	atras
realmente	poco	directamente	jamas
posteriormente	aquel	alrededor	cierto
lejos	cualquiera	aca	quizas
siquiera	aun	respectivamente	apenas
solamente	pronto	arriba	demasiado
precisamente	totalmente	vs	salvo
simplemente	ojala	donde	absolutamente
efectivamente	practicamente	aproximadamente	cuanto
recientemente	definitivamente	completamente	detras
ja	claramente	gratis	obviamente
lamentablemente	probablemente	encima	inmediatamente
rapidamente	aparte	anoche	exclusivamente
harto	abajo	particularmente	necesariamente
allende	pro	seguramente	personalmente
generalmente	afuera	igualmente	mio
suyo	fundamentalmente	analogamente	ah
oficialmente	cuando	escaso	especificamente
anteriormente	justamente	publicamente	temprano
previamente	basicamente	perfectamente	plenamente
normalmente	debajo	suyo	fuertemente
atentamente	acaso	demas	hola
triple	cuanto	expresamente	exactamente
supuestamente	analiticamente	debidamente	altamente
adicionalmente	diariamente	legalmente	delante
algun	aparentemente	paralelamente	constantemente
inicialmente	relativamente	permanentemente	adecuadamente
anualmente	adentro	excepto	facilmente
chao	profundamente	caracteristicamente	habitualmente
quiza	ampliamente	suficientemente	eventualmente
evidentemente	naturalmente	na	oportunamente
posiblemente	afortunadamente	ningun	animicamente
libremente	simultaneamente	unicamente	originalmente
significativamente	vuestro	extremadamente	formalmente
ciertamente	seriamente	proximamente	so
estrictamente	propiamente	conjuntamente	tradicionalmente
parcialmente	verdaderamente	considerablemente	sumamente
independientemente	lentamente	mayoritariamente	usted
economicamente	abiertamente	almenos	adios
desgraciadamente	quedo	positivamente	levemente
politicamente	tremendamente	historicamente	esencialmente

inclusive	dificilmente	voluntariamente	versus
correctamente	netamente	demasiado	preferentemente
tecnicamente	gravemente	indudablemente	activamente
oh	traves	curiosamente	automaticamente
mensualmente	indirectamente	duramente	sencillamente
frecuentemente	fisicamente	bastante	el
ultimamente	gratuitamente	periodicamente	temporalmente
sinceramente	integramente	tuyo	sustancialmente
antano	literalmente	favorablemente	negativamente
paulatinamente	enfrente	francamente	intensamente
sorpresivamente	precedentemente	estrechamente	radicalmente
mayormente	presuntamente	extraordinariamente	mmm
internacionalmente	sic	notablemente	chano
ay	otrora	inevitablemente	enormemente
concretamente	regularmente	socialmente	notoriamente
progresivamente	fantastico	enseguida	coercitivamente
largamente	violentamente	acreditadamente	derechamente
meramente	convenido	enteramente	demasiado
logicamente	judicialmente	cuidadosamente	ilegalmente
continuamente	presumiblemente	esto	individualmente
vos	sistematicamente	mutuamente	viceversa
categoricamente	reiteradamente	telefonicamente	gradualmente
cuan	satisfactoriamente	excesivamente	bla
potencialmente	exitosamente	virtualmente	visiblemente
brevemente	sexualmente	tranquilamente	entretanto
eh	sucesivamente	colectivamente	usualmente
explicitamente	poh	internamente	firmemente
puramente	sendos	drasticamente	ligeramente
eminentemente	tempranamente	comunmente	masivamente
felizmente	separadamente	injustamente	recien
tajantemente	profesionalmente	accionariamente	cho
cuasi	idem	prontamente	decididamente
anticipadamente	paradojicamente	contrariamente	moralmente
textualmente	geneticamente	juridicamente	puntualmente
excepcionalmente	mundialmente	severamente	eficientemente
objetivamente	antiguamente	unilateralmente	empero
brutalmente	constitucionalmente	unanimemente	responsablemente
indebidamente	eficazmente	mio	indefinidamente
extraoficialmente	repentinamente	adonde	abruptamente
eternamente	primeramente	consecuentemente	momentaneamente
fielmente	intimamente	medianamente	ocasionalmente
detalladamente	razonablemente	mentalmente	proporcionalmente
semanalmente	cabalmente	estadisticamente	verbalmente
obligatoriamente	desafortunadamente	antenoche	ostensiblemente

legitimamente	validamente	comodamente	peligrosamente
doblemente	bruscamente	administrativamente	transitoriamente
comercialmente	solidariamente	indistintamente	teoricamente
za	honestamente	rotundamente	artificialmente
anteayer	rigurosamente	democraticamente	dramaticamente
espontaneamente	fehacientemente	culturalmente	ambientalmente
insistentemente	erroneamente	desesperadamente	dignamente
irremediablemente	detenidamente	escuetamente	arbitrariamente
increiblemente	alo	sobremanera	suavemente
crecientemente	terriblemente	infinitamente	seguidamente
ups	llanamente	comparativamente	pacificamente
talvez	urgentemente	diametralmente	emocionalmente
idealmente	cientificamente	enfaticamente	escasamente
silenciosamente	ole	arduamente	tipicamente
nomas	ironicamente	profusamente	energicamente
extranamente	sicologicamente	timidamente	integralmente
tristemente	financieramente	deliberadamente	instantaneamente
indiscutiblemente	eticamente	accidentalmente	cordialmente
privadamente	simbolicamente	inesperadamente	infructuosamente
atropelladamente	aceleradamente	invariablemente	gratamente
materialmente	casualmente	apropiadamente	frontalmente
ea	equivocadamente	fo	estrategicamente
geograficamente	globalmente	maliciosamente	innecesariamente
futbolisticamente	implicitamente	manifiestamente	humanamente
ala	bah	pacientemente	coordinadamente
cuadruple	inmensamente	militarmente	doquier
milagrosamente	cotidianamente	electoralmente	fundadamente
musicalmente	adonde	convenientemente	quirurgicamente
sanamente	generosamente	validamente	obligadamente
minuciosamente	racionalmente	inexplicablemente	acertadamente
misteriosamente	provisionalmente	tragicamente	veintiuno
espiritualmente	ochocientos	tacitamente	intencionalmente
alternativamente	marcadamente	secretamente	cuan
localmente	aisladamente	popularmente	visualmente
sustantivamente	injustificadamente	terminantemente	sostenidamente
provisoriamente	reglamentariamente	tardiamente	laboralmente
carinosamente	electronicamente	prioritariamente	uy
esporadicamente	prematuramente	sorprendentemente	preliminarmente
clandestinamente	poderosamente	moderadamente	decisivamente
institucionalmente	subitamente	magistralmente	afirmativamente
incansablemente	informalmente	respetuosamente	extensamente
nitidamente	universalmente	predominantemente	repetidamente
despacio	autenticamente	uh	clinicamente
intelectualmente	tecnologicamente	ininterrumpidamente	alegremente

primordialmente	maxime	penalmente	deportivamente
funcionalmente	civilmente	supra	espectacularmente
modestamente	bum	humildemente	manualmente
sentimentalmente	intrinsecamente	inconscientemente	taxativamente
coincidentemente	tate	friamente	mecanicamente
conscientemente	inteligentemente	matematicamente	otro
territorialmente	quienquiera	substancialmente	equitativamente
extrajudicialmente	incondicionalmente	apud	impecablemente
sutilmente	descaradamente	estrepitosamente	finamente
merecidamente	sensiblemente	impunemente	setecientos
armonicamente	discretamente	religiosamente	inexorablemente
amablemente	criticamente	tenazmente	raramente
reciprocamente	precozmente	estructuralmente	trimestralmente
forzosamente	joder	complementariamente	solemnemente
velozmente	cronologicamente	artisticamente	mortalmente
quimicamente	celosamente	externamente	involuntariamente
conceptualmente	ergo	exhaustivamente	consistentemente
genericamente	apresuradamente	inapelablemente	primitivamente
semestralmente	gentilmente	ordenadamente	ciegamente
determinadamente	indiscriminadamente	salvajemente	veintitres
exageradamente	densamente	ideologicamente	tacticamente
ordinariamente	esteticamente	minimamente	interinamente
afanosamente	cualitativamente	discrecionalmente	transversalmente
digitalmente	falsamente	imperiosamente	solidamente
subsidiariamente	sumariamente	centralmente	raudamente
horizontalmente	remotamente	habilmente	veintiuno
inequivocamente	ferreamente	inutilmente	sigilosamente
interiormente	holgadamente	verticalmente	marginalmente
sabiamente	forte	intempestivamente	consiguientemente
prudencialmente	lealmente	afectuosamente	consecutivamente
ibidem	someramente	agresivamente	tangencialmente
criminalmente	irresponsablemente	brillantemente	asimismo
genuinamente	sagradamente	idem	amargamente
irregularmente	graficamente	soberanamente	psicologicamente
indefectiblemente	retroactivamente	sospechosamente	tentativamente
ingenuamente	groseramente	reconocidamente	sobradamente
supletoriamente	creativamente	resueltamente	autonomamente
incesantemente	cuantitativamente	vagamente	empiricamente
cerquita	elegantemente	ilicitamente	medicamente
entusiastamente	malamente	superficialmente	fisicamente
maravillosamente	ejem	insuficientemente	inusualmente
experimentalmente	fatalmente	copulativamente	latamente
hipoteticamente	limpiamente	vertiginosamente	indisolublemente
pasivamente	apasionadamente	circularmente	inversamente

contractualmente	anonimamente	mancomunadamente	providencialmente
animadamente	certeramente	desinteresadamente	livianamente
alternadamente	dolorosamente	estoicamente	turisticamente
presuntivamente	regionalmente	ui	biologicamente
convencionalmente	cruelmente	irrevocablemente	ecologicamente
fijamente	reservadamente	juntamente	uniformemente
meridianamente	airadamente	despacito	despectivamente
persistentemente	ax	cobardemente	majaderamente
preventivamente	prolijamente	ocupacionalmente	corrientemente
dolosamente	orgullosamente	inocentemente	policialmente
acuciosamente	placidamente	defensivamente	justificadamente
magicamente	vigorosamente	compulsivamente	originariamente
primariamente	computacionalmente	ineludiblemente	grandemente
organicamente	torpemente	veladamente	ferozmente
imperativamente	jerarquicamente	debilmente	porcentualmente
abundantemente	centralizadamente	honradamente	quia
vanamente	hondamente	perdidamente	acidamente
fugazmente	fluidamente	coherentemente	fraudulentamente
valientemente	selectivamente	recurrentemente	competitivamente
previsiblemente	subrepticiamente	pobremente	procesalmente
retrospectivamente	academicamente	amistosamente	electricamente
postumamente	fervientemente	imprudentemente	perpetuamente
extramuros	fraternalmente	precipitadamente	instintivamente
privativamente	todito	vivamente	correlativamente
magnificamente	pesadamente	erradamente	forzadamente
mal	nominalmente	operativamente	pausadamente
singularmente	horriblemente	quintuple	aleatoriamente
excelentemente	hermeticamente	oralmente	asa
eufemisticamente	impostergablemente	perentoriamente	ruidosamente
espacialmente	ibidem	metodologicamente	premeditadamente
diplomaticamente	lastimosamente	penosamente	graciosamente
intencionadamente	afectivamente	estupendamente	imprevistamente
periodicamente	optimamente	nerviosamente	secundariamente
artesanalmente	poeticamente	disciplinariamente	hidalgamente
incorrectamente	chus	literariamente	vastamente
abrumadoramente	desfavorablemente	dondequiera	olimpicamente
dulcemente	irreversiblemente	linealmente	manosamente
metaforicamente	ofensivamente	vulgarmente	bilateralmente
delicadamente	crudamente	lateralmente	numericamente
molto	probadamente	heroicamente	locamente
mondo	exteriormente	increiblemente	monetariamente
alfabeticamente	cercanamente	circunstancialmente	restrictivamente
tontamente	condicionalmente	nacionalmente	serenamente
obsesivamente	exponencialmente	medio	tematicamente

contradictoriamente	familiarmente	organizadamente	prudentemente
ansiosamente	intensivamente	vivace	implacablemente
industrialmente	longitudinalmente	paf	bellamente
operacionalmente	peyorativamente	passim	asombrosamente
desordenadamente	asiduamente	freneticamente	homogeneamente
sucintamente	tibiamente	denodadamente	inadvertidamente
sniff	subjetivamente	explosivamente	calladamente
concienzudamente	equilibradamente	ahorita	eureka
admirablemente	inadecuadamente	intuitivamente	subterraneamente
abusivamente	disimuladamente	mezzo	ajustadamente
ancestralmente	decentemente	efusivamente	productivamente
incuestionablemente	innegablemente	que	relajadamente
disciplinadamente	meticulosamente	preferencialmente	clasicamente
ejemplarmente	filosoficamente	ulteriormente	incidentalmente
interminablemente	ox	armoniosamente	cautelosamente
ejecutivamente	imaginariamente	otrosi	porfiadamente
circunstanciadamente	contundentemente	descarnadamente	diligentemente
quincenalmente	temerariamente	solapadamente	tiernamente
vilmente	bulliciosamente	encarecidamente	intermitentemente
escrupulosamente	amigablemente	calurosamente	infraganti
lueguito	precariamente	rutinariamente	agilmente
constructivamente	copiosamente	desconsoladamente	despiadadamente
metodicamente	unitariamente	amorosamente	acriticamente
desproporcionadamente	ilegitimamente	pormenorizadamente	transparentemente
vitalmente	calmadamente	doctrinariamente	pesimamente
periodisticamente	piu	caprichosamente	desmesuradamente
irremisiblemente	resumidamente	estupidamente	insoportablemente
sinteticamente	exclusive	ricamente	sanitariamente
concertadamente	infortunadamente	absurdamente	angustiosamente
arquitectonicamente	estacionalmente	gustosamente	fisiologicamente
pretendidamente	agudamente	anormalmente	astutamente
clo	estimativamente	imperceptiblemente	intramuros
irrestrictamente	ridiculamente	zas	hermosamente
huifa	secuencialmente	agonicamente	convincentemente
irracionalmente	reflexivamente	burdamente	entranablemente
escandalosamente	insolitamente	mansamente	tributariamente
dinamicamente	esplendidamente	impensadamente	trabajosamente
deficientemente	sobriamente	poco	cortesmente
prestamente	avanti	coyunturalmente	documentalmente
propio	verazmente	acabadamente	agradablemente
fortuitamente	rubato	solitariamente	contablemente
etimologicamente	logisticamente	patrimonialmente	fieramente
opcionalmente	apretadamente	civilizadamente	estatutariamente
rectamente	cronicamente	furiosamente	impresionantemente

nominativamente	pateticamente	teatralmente	obstinadamente
pomposamente	descuidadamente	laconicamente	verbigracia
visceralmente	adrede	calidamente	comprensiblemente
desmedidamente	dificultosamente	fiscalmente	interesadamente
similarmente	vergonzosamente	acaloradamente	caballerosamente
calculadamente	deliciosamente	legislativamente	pecuniariamente
pedagogicamente	tenuemente	desigualmente	elocuentemente
preponderantemente	secamente	sectorialmente	sensualmente
esquematicamente	imparcialmente	soterradamente	fallidamente
rigidamente	tenisticamente	adversamente	atinadamente
confidencialmente	conocidamente	despreocupadamente	generacionalmente
geometricamente	ingeniosamente	dialecticamente	fervorosamente
furtivamente	febrilmente	flagrantemente	huy
importantemente	morfologicamente	simetricamente	aparatosamente
audazmente	cinicamente	contemporaneamente	dudosamente
exquisitamente	ficticiamente	incomparablemente	inusitadamente
junto	laboriosamente	lejanamente	siquicamente
vehementemente	amenamente	confiadamente	corporalmente
didacticamente	enganosamente	jocosamente	miserablemente
ritmicamente	sociologicamente	teologicamente	ventajosamente
anatomicamente	buenamente	deslealmente	hipocritamente
indiscutidamente	plasticamente	psiquicamente	triunfalmente
animosamente	cumplidamente	emotivamente	documentadamente
descontroladamente	linguisticamente	atrozmente	proactivamente
publicitariamente	etnicamente	opticamente	acuciantemente
corporativamente	cristianamente	distraidamente	histologicamente
infundadamente	sustentablemente	urbanisticamente	ciclicamente
correspondientemente	doceavo	energeticamente	esforzadamente
irreparablemente	sextuple		

Fuente: Elaboración a partir de [67] y http://www.ranks.nl/stopwords/spanish
Tabla B.1: Lista de Stop Words usadas en español