Una aplicación de Web Opinion Mining para la extracción de tendencias y tópicos de relevancia a partir de las opiniones consignadas en blogs y sitios de noticias

Rodrigo Dueñas F. [*]
Juan D. Velásquez[*]

Resumen

El análisis de tendencias se ha abordado tradicionalmente a través de la realización de encuestas, las cuales poseen un alto contenido de subjetividad y las respuestas se ven constantemente afectadas por factores exógenos al evento bajo estudio. Este exceso de factores exógenos y subjetividad puede conducir a errores significativos, basta con ver los resultados de la última encuesta para la elección de alcaldes, la que predijo de manera errónea qué candidatos ganarían en las comunas mas emblemáticas de Santiago. En este trabajo, presentamos una metodología alternativa para detectar tendencias en la Web, a través del uso de técnicas de recuperación de la información, modelamiento de tópicos y minado de opiniones. Dado un conjunto de sitios semilla, se procede a extraer los tópicos que se mencionan en los documentos recuperados desde ellos y posteriormente se acude a las redes sociales para obtener la opinión por parte de sus usuarios en relación a estos. Usando esta metodología de detección de tendencias es posible complementar la información extraída a través de metodologías tradicionales para predecir eventos y reducir los efectos de los factores exógenos introducidos por los medios tradicionales.

Palabras Clave: *Opiniones*, *Tendencias*, *Web Opinion Mining*, *Tópicos*, *Blogs*, *Noticias*.

[*]Departamento de Ingeniería Industrial, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile, Santiago, Chile.

1. Introducción

Con la aparición de las aplicaciones web que permiten la creación de contenido y la colaboración por parte de los usuarios, como lo son *wikis* y *blogs*, (las cuales darían el puntapié inicial a lo que ahora se conoce como la Web 2.0) la función de la Web en la sociedad mundial se vio fuertemente potenciada. Esta dejó de ser tan sólo un repositorio de información y se transformó en un canal interactivo entre todas las entidades que la componen, tanto usuarios como proveedores de información. Este cambio de paradigma permitió que todos ellos pudiesen contribuir activamente a la creación de contenido, provocando un explosivo aumento en la participación de sus usuarios en la Web, y por consiguiente de la cantidad de información y conocimiento disponible en ella.

Junto a ello, Internet trajo consigo cambios drásticos en la manera que se interactúa en el mercado empresarial. Estos cambios provocan que una empresa pueda crecer en muchas direcciones y no sólo aumentando la cantidad de productos que produce o el número de personas a las que ofrece sus servicios. Así, una vez que una empresa decide crecer, ya sea expandiendo negocios hacia nuevos mercados u ofreciendo nuevos productos y servicios, la cantidad de información externa que debe abarcar para poder realizar una buena toma de decisiones se vuelve cada vez mayor, por lo que debe analizar un conjunto siempre creciente de fuentes de información para poder recuperar el conocimiento necesario para que este análisis sea valioso para la empresa.

En esta misma línea, es cada vez más necesario ser capaz de manejar grandes volúmenes de datos para gestionar de la mejor manera posible los recursos que se disponen, y al mismo tiempo anticiparse a cada movimiento que realizará la competencia en busca de obtener ventajas competitivas, o impedir que otros las obtengan, para ser líderes en el mercado. El primer problema al que se debe enfrentar una empresa sumergida en el mundo globalizado, es el más complejo desde el punto de vista de la gestión de operaciones, por lo que varias metodologías y herramientas han nacido proponiendo soluciones, entre las cuales se encuentran los *Data Warehouses*, la *Business Intelligence* y el recientemente acuñado término de *BigData*. El segundo problema no sólo involucra a la gestión de operaciones, ya que es necesario tener un equipo multidisciplinario encargado constantemente de monitorear el mercado, las acciones de las otras empresas, los anuncios presentes en los medios y cualquier indicio que permita anticiparse a los lanzamientos de productos y servicios de la competencia.

Una posible solución al problema planteado es minar la web en busca de

esos indicios de manera automática, con foco en qué tópicos se habla en la web, y analizando las redes sociales para estimar que percepción poseen los cibernautas sobre ellos. Es factible plantear la hipótesis de que a través de realizar un análisis de gran parte del conocimiento objetivo generado por los usuarios y los medios se puede atisbar aquellos indicios claves a la hora de plantear una planificación estratégica y operacional. Un sistema capaz de realizar esto de manera aproximada es realizable utilizando técnicas de recuperación de la información, modelamiento de tópicos y minado de opiniones sobre fuentes cuidadosamente seleccionadas que sean capaces de otorgarle al sistema una muestra significativa de todo lo que se habla sobre los mercados en los cuales se ve inmersa la empresa a nivel competitivo.

Por lo mencionado anteriormente, se plantea como hipótesis de investigación que es posible extraer tendencias y obtener una representación aproximada del comportamiento de estas a través del análisis de los documentos presentados en sitios de noticias y las opiniones consignadas en las redes sociales por parte de sus usuarios.

En la segunda sección de este artículo se da a conocer el estado del arte respecto de las técnicas de recuperación de información, modelamiento de tópicos en documentos y finalmente sobre algoritmos de minado de opiniones. En la sección 3, se da a conocer en detalle el modelo propuesto para la detección de tendencias en la Web, en particular la detección de tópicos y el minado de opiniones referentes a estos, los cuales serán evaluados a través de los experimentos presentes en la sección 4. Para finalizar, en la quinta sección de este artículo se presentan las conclusiones relevantes al trabajo desarrollado y posibles ramas de investigación a futuro.

2. Trabajo relacionado

2.1. Modelos de Tópicos

Un modelo de tópicos tiene como objetivo identificar las relaciones latentes entre documentos pertenecientes a una colección, con el fin de dar una descripción sucinta de esta sin perder información desde el punto de vista estadístico.

El precursor de los modelos de tópicos es David Blei, el cual en [4] describe de manera detallada los modelos de tópicos y las aplicaciones de estos. En ella, se define un tópico como el conjunto de elementos que pueden representar una temática presente en una colección de documentos sin pérdida de información estadística. Por ejemplo, si existe una colección de documentos textuales

que abarca múltiples temas, un tópico es un conjunto de palabras que logra describir estadísticamente uno de estos temas.

Entre los modelos de tópicos existente, los más utilizados son los desarrollados por Blei *et al.* Entre ellos, los más populares son el modelo LDA (Latent Dirichlet Allocation) [4] y el modelo CTM (Correlated Topic Model) [3].

Estos modelos de tópicos se cimentan sobre las siguientes definiciones:

- Una *palabra* w es la unidad elemental de un documento textual y se define como un elemento de un vocabulario indexado V. Para efectos de estos modelos, para representar una palabra se hace uso de vectores unitarios en donde la n-ésima palabra de V se representa con un vector de largo $|V|$ en el cual sólo su componente n-ésima es igual a 1.
- Un *documento* es un arreglo de palabras descrito como $\mathbf{w} = (w_1, w_2, ..., w_N)$, donde w_n es la n-ésima palabra de este.
- Un *corpus* es una colección de documentos descrita como $D = \{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_M\}$.
- Un *tópico* es una distribución de probabilidad sobre un vocabulario V fijo. Por ejemplo, el tópico *política* está descrito por palabras como *partido, diputado, senado, ley* de manera frecuente y palabras como *guerra, marcador, gol* con probabilidad casi nula.

A continuación se da a conocer una descripción de cada uno de los modelos mencionados anteriormente, dando a conocer las diferencias entre estos y las principales características de cada uno de ellos.

2.1.1. Latent Dirichlet Allocation

El modelo llamado *Latent Dirichlet Allocation*[4] es considerado el más sencillo de los modelos de tópicos presentes hoy en día, y por ello es utilizado frecuentemente en aplicaciones que requieran obtener información sobre colecciones de documentos de manera rápida y eficiente.

El modelo LDA trabaja bajo el supuesto de que los tópicos presentes en la colección de documentos que se está analizando no necesariamente están relacionados y por consiguiente no dependen entre ellos.

Para extraer la estructura de tópicos presente en una colección, este modelo hace uso de un modelo estadístico de generación de documentos, tópicos y palabras a lo largo del tiempo que abarque esta. El siguiente proceso se realiza para cada documento presente en una colección:

1. Definir una distribución aleatoria para la presencia de los tópicos en la colección y una distribución para la presencia de las palabras para cada tópico que se desea encontrar.

2. Luego, por cada palabra presente en el documento bajo análisis se debe:

   a) Escoger un tópico aleatoriamente haciendo uso de la distribución generada en el paso 1.

   b) Escoger una palabra del documento aleatoriamente a partir de la distribución del vocabulario en relación al tópico escogido.

Formalmente, para determinar la estructura de tópicos existente luego del proceso de generación, es necesario calcular las distribuciones condicionales entre los tópicos y sus documentos, el cual es un problema NP completo debido a que la cantidad de estructuras que pueden representar una colección de documentos crece exponencialmente en relación a la cantidad de documentos y palabras presente en esta. Este proceso es descrito formalmente como sigue:

1. Escoger $N \sim Poisson(\xi)$
2. Escoger $\theta \sim Dirichlet(\alpha)$
3. Para cada palabra w_n en $\mathbf{w}$

   a) Escoger un tópico $z_d \sim Multinomial(\theta)$

   b) Escoger una palabra w_d a partir de $p(w_n|z_n, \beta)$, la distribución multinomial de probabilidades condicionada sobre el tópico z_n.

Donde cada variable del proceso corresponde a:

- β es la matriz de probabilística de que el documento contenga la palabra w^j dado que discute el tópico z^i, con $B_{ij} = p(w^j = 1|z^i = 1)$.
- θ_d es la distribución de tópicos para el documento d, es decir, el conjunto de probabilidades $\theta_{d,k}$ donde esta corresponde a la probabilidad de que el documento d trate del tópico k.
- z_d son las asociaciones de tópicos para el documento d con $z_{d,n}$ es el tópico asociado a la palabra n-ésima del documento d
- w_d es el conjunto de palabras presentes en el documento d.
- $w_{d,n}$ es la palabra n-ésima del documento d.

A partir de esto, es posible definir el proceso generativo de documentos a través de la distribución conjunta de variables observables y no observables como se define en la ecuación 1. La solución a esta ecuación puede ser obtenida haciendo uso de algoritmos de inferencia estadística como el algoritmo *Sampleo*

de Gibbs, los que además de estimar la estructura de tópicos de una colección, permiten inferir la estructura de tópicos presente en otros *corpus* que estén compuestos de documentos que hablen de temas similares a los utilizados para entrenar el modelo.

$$p\left(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}\right) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n|\beta_{1:K}, z_{d,n}}) \right) \tag{1}$$

2.2. Modelos de extracción de opiniones

Con el nacimiento de las redes sociales y la llegada de la Web 2.0, los usuarios comenzaron ser capaces de generar nuevo contenido en la web y también de dar a conocer sus opiniones sobre variados hechos, productos, servicios y cualquier otro tema que sea susceptible de generar un sentimiento o una opinión en ellos.

Para aprovechar esta nuea información que está siendo generada en la web se han desarrollado una serie de metodologías, algoritmos y técnicas para recuperar información desde documentos opinados. Esta nueva rama de la recuperación de la información es llamada *Web Opinion Mining*, la cual tiene como objetivo principal extraer información a partir de las opiniones que se encuentran en los documentos opinados publicados en la web [9].

Los modelos de opinión son utilizados frecuentemente en donde es necesario hacer uso de las opiniones de los usuarios para evaluar u obtener información sobre productos y servicios. En [15] se menciona que los algoritmos de minado de opiniones son utilizados frecuentemente en detección de spam en review de productos, creación y mejoramiento de sistemas de recomendación de productos y servicios o de avisaje online, evaluación de nuevos productos en la web, evaluar el impacto que tienen las reviews en las utilidades de un negocio o un producto, etc.

Una opinión se define como una creencia subjetiva por parte de un sujeto sobre algún objeto, tema o situación en particular, que nace de una interpretación emocional por parte de éste del objeto bajo análisis o una característica de este [6]. Por consiguiente, una opinión es una creencia subjetiva de un *emisor* sobre el *receptor* o una característica de este, y posee una polaridad que señala el tipo de emoción (positiva o negativa) que da paso a la opinión propiamente tal.

Los modelos de extracción o minado de opiniones trabajan sobre *documentos opinados*, los cuales son definidos en [9] como todo documento que contenga una o más oraciones que expresan una opinión. Por lo tanto, se puede decir que

los modelos de extracción de opiniones buscan determinar qué tipo de emoción motiva la emisión de una opinión en un documento [6] o que polaridad es la predominante en este [18].

Para dar a conocer un modelo de opiniones es necesario presentar una serie de definiciones que dan sustento a la gran mayoría de los modelos utilizados en la actualidad. Estas definiciones son las que siguen:

- **Objeto**: Un *objeto* o es una entidad que puede ser un producto, un servicio, un individuo, una organización, un evento, etc. descrito por la dupla (T, A) donde T es la jerarquía que describe cada una de las componentes del objeto y A es el conjunto de atributos de este. A su vez, cada componente posee su propio conjunto de sub-componentes y atributos.

- **Opinión**: Una *opinión* sobre una característica f objeto o es una evaluación emocional que realiza un *emisor* sobre este o una característica de él.

- **Emisor**: El *emisor* de una opinión es aquella persona u organización que la expresa.

- **Polaridad**: La *polaridad* de una opinión indica si la opinión es *positiva*, *negativa* u *objetiva*.

Además, en el modelo de análisis de opiniones basado en características, un objeto o se describe como un conjunto de características $F = f_1, f_2, ..., f_n$ donde también se incluye el objeto en cuestión como una característica particular. En este caso, cada característica f_i puede ser descrita por el conjunto de palabras o frases $W_i = w_{i1}, w_{i2}, ..., w_{im}$, donde cada w_{ij} es un sinónimo de la característica f_i; además, f_i también puede ser expresada a través del conjunto de indicadores de característica $I_i = i_{i1}, i_{i2}, ..., i_{iq}$.

Bajo este modelo, un documento d que contiene opiniones es descrito como aquel que contiene opiniones sobre un conjunto de objetos $o_1, o_2, ..., o_q$ emitidas por un conjunto de emisores $h_1, h_2, ..., h_p$. En este caso, cada opinión o_j se enfocan en un subconjunto F_j de características del objeto en cuestión y puede ser clasificada en uno de los siguientes tipos:

- **Opinión directa**: Es la quintupla $(o_j, f_{jk}, oo_{ijkl}, h_i, t_l)$, donde o_j es el objeto sobre el cuál consiste la opinión, f_{jk} es la característica del objeto o_j que está siendo analizada, oo_{ijkl} es la polaridad de la opinión sobre la característica f_{jk}, h_i es el emisor de la opinión y finalmente, t_l es el momento en el cuál h_i expresó la opinión.

- **Opinión comparativa**: Expresa la relación, sea esta de similitud o de diferencia entre dos o más objetos y las preferencias del emisor de la opinión sobre un conjunto común de características entre los objetos.

Toda opinión se basa en las emociones que guían al emisor a emitirla en el momento que este acto sucede. De acuerdo a lo expresado en [9], las emociones son ***sentimientos y pensamientos subjetivos***, y estas se dividen en 6 tipos primarios: *amor, alegría, sorpresa, rabia, tristeza* y *temor*.

Si bien todas las opiniones nacen de una emoción, la manera en que estas son expresadas por el emisor de ellas permite clasificarlas en dos tipos: las opiniones *explícitas*, aquellas en que el emisor expresa claramente la opinión a través de una frase subjetiva; y las *implícitas*, donde la opinión en cuestión es expresada a través del uso de una frase objetiva. Un ejemplo de opinión explícita es *"me encanta el sabor de este helado"* y de opinión implícita es *"la linterna explotó a la semana de haberla comprado"*.

2.2.1. Aplicaciones de los algoritmos de minado de opiniones

Entre las aplicaciones que tienen los algoritmos de minado de opiniones podemos encontrar:

1. **Análisis de reviews de productos**: En [2] se discuten distintas aplicaciones de estos algoritmos en el análisis de reviews de productos, entre ellas se destacan el resumen de opiniones, detectar reviews falsos o spam y la evaluación monetaria de las características de un producto.

2. **Sistemas de recomendación**: En [7] se estudia mejorar sistemas de recomendación de productos a través del uso de las opiniones emitidas por usuarios de estos mismos sistemas.

3. **Política**: En [13] se muestran distintos enfoques para analizar campañas políticas y la percepción de la gente sobre leyes y candidatos políticos.

2.2.2. Algoritmos para extracción de polaridad de opiniones

En la plataforma de detección de tendencias se hace uso de algoritmos de detección de polaridad para determinar qué es lo que se opina en la web sobre los tópicos que son extraídos desde los documentos recuperados. A continuación se dará a conocer las dos afluentes más utilizadas de algoritmos de detección de polaridad en opiniones.

Algoritmos de clasificación a través de aprendizaje supervisado: La mayoría de los algoritmos de aprendizaje supervisado existentes (Naive-

Bayes, Support Vector machines, etc.) pueden ser aplicados a la clasificación de polaridad de documentos tal como se muestra en [16, 11].

El algoritmo más utilizado debido a su simplicidad es un clasificador Naive-Bayes, el cual busca obtener las probabilidades de que un documento d posea la polaridad p $\Pr(p \mid d)$. Este tipo de clasificador obtiene estas probabilidades al resolver el siguiente problema de maximización: $arg \max_{p \in P}\{\Pr(p \mid d)\}$.

Los clasificadores de Naive-Bayes hacen uso de la regla de Bayes para poder simplificar el problema de maximización que deben resolver:

$$p_d = arg \max_{p \in P} \left\{ \frac{\Pr(d \mid p) \cdot \Pr(p)}{\Pr(d)} \right\} \tag{2}$$

Debido a que sólo se busca conocer la probabilidad de que un documento tenga una polaridad y no obtener un puntaje específico para el nivel de polaridad que posee, el denominador de la ecuación 2 puede ser eliminado. Esto junto con el hecho de que uno de los supuestos del clasificador de Naive-Bayes es la independencia condicional entre todas las polaridades, se puede decir que:

$$\Pr(d \mid p) = \prod_{i=1}^{m} \Pr(w_i \mid p) = \prod_{i=1}^{m} \frac{\#(w_i, p)}{\#(w_i)} \tag{3}$$

Con $\#(w_i, p)$ el número de veces que la palabra w_i se ha encontrado en documentos de polaridad p en el conjunto de entrenamiento y $\#(w_i)$ el número de veces que la palabra w_i aparece en este último. Para evitar que existan probabilidades 0, se realiza un proceso llamado "suavización de Laplace"que consiste en lo siguiente:

$$\Pr(d \mid p) = \prod_{i=1}^{m} \frac{\#(w_i, p) + 1}{\#(w_i) + m} \tag{4}$$

Con estas ecuaciones basta resolver el problema de maximización planteado para obtener las probabilidades de que cada documento posea una polaridad en particular.

En general, las características utilizadas por los algoritmos de aprendizaje supervisado se dividen en las siguientes categorías:

- *Frecuencia y presencia de términos*: Si bien el uso de frecuencia de aparición de términos, por ejemplo a través del modelo *tf-idf*, en la recuperación de la información siempre ha sido de mucha utilidad, en [16] se muestra que en el caso de la extracción de opiniones desde documentos la *presencia* de un término es más importante que la frecuencia con que este aparece.

- *Partes del discurso*: Los adjetivos han sido utilizados con frecuencia [11] en el uso de algoritmos de aprendizaje supervisado ya que existe una alta correlación entre la presencia de adjetivos en una oración y la subjetividad de esta.
- *Sintáxis*: En [12] se hace uso de la relación entre las palabras como características en algoritmos de aprendizaje supervisado.

Algoritmos de clasificación a través de aprendizaje no supervisado: En [19] se propone un algoritmo de aprendizaje no supervisado para la clasificación de polaridad de documentos que se compone de tres etapas:

1. Se extraen todas las frases con verbos o adjetivos, ya que tal como se menciona en [11], estas partes del discurso se han mostrado muy útiles a la hora de detectar opiniones en documentos. Sin embargo, a pesar de que un adjetivo por si solo puede demostrar subjetividad, puede que no exista la información suficiente para determinar la polaridad de la opinión. Debido a esto, este algoritmo trabaja con pares de palabras, una de ellas siendo un adjetivo y la otra una palabra contextual que facilita la determinación de la polaridad de la oración en cuestión. Estos pares de palabras son extraídos siempre y cuando, considerando las dos palabras y la que les sigue, correspondan a alguno de los patrones conocidos.
2. Se estima la polaridad de las frases extraídas, haciendo uso de la métrica de dependencia estadística entre términos llamada *pointwise mutual information* (PMI) que se presenta en la ecuación 5

$$PMI(w_1, w_2) = log_2 \left(\frac{\Pr(w_1 \wedge w_2)}{\Pr(w_1)\Pr(w_2)} \right) \tag{5}$$

   Luego, la polaridad de una frase puede ser calculada basándose en el nivel de asociación entre ella y las palabras de referencia *pobre* y *excelente* a través de la ecuación 6

$$oo(frase) = PMI(frase, "excelente") - PMI(frase, "pobre") \tag{6}$$

3. Finalmente, el algoritmo calcula la polaridad *oo* promedio de todas las frases en el documento y lo clasifica dependiendo de si el promedio es positivo o negativo.

Algoritmos basados en lexicones de opinión: Los algoritmos basados en lexicones de opinión son los algoritmos más sencillos y a su vez los que buscan ser de uso más general debido a que la información utilizada para determinar la polaridad de una opinión no está restringida a ningún dominio en particular. Un lexicón es un conjunto de palabras rotuladas con polaridad de sentimientos, es decir, cada palabra perteneciente al lexicón tiene asociado un puntaje de polaridad.

Estos algoritmos trabajan bajo la hipótesis de que una palabra es considerada la unidad elemental de una opinión y por lo tanto la polaridad de una opinión puede reconstruirse a partir de la polaridad de cada una de las palabras que la componen. Ejemplos de algoritmos que hacen uso de lexicones para determinar la polaridad de una opinión se pueden encontrar en [14, 16]. En relación al minado de opiniones desde documentos de microblogging, Kouloumpis et al. dan a conocer en [8] que los algoritmos de basados en lexicones pueden dar buenos resultados.

El lexicón utilizado por la plataforma de detección de tendencias es *SentiWordNet* el cual está disponible públicamente para ser usado en este tipo de aplicaciones de minado de opiniones.

Cada palabra presente en un lexicón tiene asociado un puntaje por cada polaridad positiva, negativa y objetiva que representan el aporte de esta palabra para la polaridad de una opinión. En el caso de *SentiWordNet* [14] se tiene que cada palabra tiene asociado sólo los puntajes de polaridad positiva w^p y negativa w^n, y además el puntaje de objetividad $w^o = 1 - w^p + w^n$.

Los algoritmos basados en lexicones de opinión hacen uso de las siguientes metodologías para reconstruir la polaridad de la opinión contenida en un documento a partir de sus palabras:

- **Conteo de palabras**: los puntajes de polaridad de un documento se obtiene a través de la fracción de palabras cuya que posee una polaridad predominante p. En este caso, una palabra será considerada de una polaridad p si su mayor puntaje es el de aquella polaridad.

- **Promedio de palabras**: En un algoritmo de promedio de palabras, el puntaje asociado a una polaridad p es el promedio de los valores de polaridad p de todas las palabras presentes en el documento.

A partir de estas metodologías básicas se pueden realizar diversas variaciones tales como: modificar los puntajes de cada palabra en base al conjunto de palabras que la rodean en el documento, hacer uso de las negaciones y la capitalización, y finalmente incorporar al puntaje la existencia de intensificadores y disminuidores de adjetivos.

2.3. Modelos de detección de Tendencias

Los modelos de detección de tendencias buscan modelar el comportamiento de los tópicos tanto desde el punto de la cobertura que este tenga en la Web, como también la percepción que los usuarios de esta tengan sobre él. Por esto, los modelos de detección de tendencias van un paso más allá que los modelos de *topic tracking*, buscando analizar también la percepción que tiene la sociedad del tópico en particular y en cómo ambas componentes se relacionan para convertir un tópico en una tendencia.

En los últimos años se han realizado variados acercamientos a la detección de tendencias en la Web. Aplicaciones enfocadas en el uso de *key-words* para extraer tendencias en Web-usage mining se presentan en [21], política [13], finanzas [17] y sistemas de recomendación [7]. Sin embargo, la contribución de este trabajo se asemeja más a metodologías genéricas que van más allá de un dominio en particular, como la presentada en [20], cuyo foco es principalmente el cómo construir una plataforma de detección de tendencias sobre una arquitectura de *cloud computing*, y no en como recuperar la información necesaria ni como decidir si es que un tópico discutido en un conjunto de documentos a lo largo del tiempo refleja una tendencia.

3. Detección de Tendencias en la Web

En la actualidad, el análisis de tendencias se ha abordado tradicionalmente a partir de encuestas, las cuales poseen un alto contenido de subjetividad, y puede conducir a errores significativos a la hora de representar los hechos que sucederán en el futuro. Estos errores pueden darse debido al contexto en que las encuestas son realizadas, las motivaciones que la gente tiene a la hora de responder y otros factores exógenos al instrumento en si.

Por otro lado, en las redes sociales, las opiniones consignadas por sus usuarios son una expresión neta de sus sentimientos. Al ser estas no obligadas ni apresuradas, es posible complementar los resultados obtenidos a través de las encuestas con un análisis de estas, reduciendo el ruido producido debido a los factores previamente mencionados.

Para comprobar la hipótesis mencionada en la primera sección de este artículo, se diseñó una plataforma de detección de tendencias que analiza la información presente en la Web en dos ejes: el primero se enfoca en el análisis de eventos, que viene dado por los documentos presentes en los sitios de noticias; y aquel que trata de los sentimientos que expresan los usuarios de las redes sociales sobre aquellos eventos. Cabe destacar que esta plataforma hace uso de

técnicas existentes de algoritmos de recuperación de la información y también modelos de tópicos y minado de opiniones para modelar cómo los tópicos se comportan a lo largo del tiempo en busca de identificar tendencias en la Web.

Inicialmente se describirá el enfoque utilizado para recuperar noticias y extraer qué tópicos están siendo discutidos en la blogosfera y en los sitios de noticias. Posteriormente se dará a conocer la metodología para extraer las opiniones a partir de los documentos publicados por los usuarios de las redes sociales, y finalmente la metodología utilizada para juntar ambos conjuntos de información con el fin de identificar tendencias en la Web.

3.1. Plataforma de Detección de Tendencias

Tal como se menciona en la sección de trabajo relacionado, los modelos de tendencias no sólo buscan detectar qué tópicos se discuten a lo largo del tiempo en un corpus de documentos, también tienen como objetivo analizar las reacciones sociales que provocan estos tópicos . En el caso de este trabajo de investigación, se acotan a las opiniones consignadas por los usuarios de redes sociales a lo largo del periodo de análisis. Así, la plataforma propuesta debe estar formada por dos pilares fundamentales: la detección de tópicos a lo largo del tiempo, y el análisis de dichos tópicos en las redes sociales a través del minado de las opiniones que les competen.

3.1.1. Minado de noticias

Se considera que una fuente de documentos presente en la web es un *feed* si cada elemento que esta contenga es desplegado de manera cronológica y pertenecen todos una misma temática. Si una fuente de documentos dispone de un punto de acceso donde se puedan recuperar cada uno de los documentos existentes en ella se dice que es un *feed sindicable*, un ejemplo de esto son todos aquellos sitios web que tienen la opción de suscribirse a su contenido a través de RSS.

Una limitante a considerar a la hora de trabajar con feeds sindicables es que el conjunto de documentos presentes cuando se accede a esta depende del tiempo. Esto implica que el conjunto de documentos $\{d_i^F\}_{i\in\mathbb{N}}$ que se obtienen al solicitar todos los documentos desde la fuente F se ve limitada por el momento t en el cual se realice esta petición. En este caso, se define $\{d_i^{F^t}\}_{i\in\mathbb{N}}$ como el conjunto de documentos recuperados desde una fuente F en un instante de tiempo t. Además, se define $\{F_i\}_{i\in\mathbb{N}}$ como el conjunto de *feeds* que recorrerá el módulo de recuperación de documentos a través de su *crawler* para alimentar el módulo de extracción de tópicos.

Para este proyecto, sólo se trabajará con feeds sindicables, por lo que, en

base a lo anterior es posible definir un algoritmo de recuperación de documentos a partir de una lista de fuentes $\{F_i\}_{i\in\mathbb{N}}$ sindicables (sean estas *RSS* o *Atom*) como se describe en el algoritmo 3.1:

Algoritmo 3.1: Recuperación de documentos

Data: $\{F_i\}_{i\in\mathbb{N}}, t$
Result: $\bigcup_i \{d_j^{F_i^t}\}_{j\in\mathbb{N}}$

```
1 documents := [];
2 for i ← 1 to ‖{F_i}_{i∈ℕ}‖ do
3 |   document ← retrieveDocument(F_i);
4 |   documents ← documents ⋃ document;
5 return documents;
```

Para extraer qué tópicos se discuten a lo largo del tiempo en la Web, se propone utilizar un enfoque basado en modelo de tópicos, debido a que permiten de manera directa obtener las *keywords* necesarias para posteriormente extraer las opiniones presentes en las redes sociales, y además, permiten monitorear la evolución de los tópicos a lo largo del tiempo. El hacer uso de técnicas de *topic tracking* o *topic detection* no es recomendable debido a las limitaciones que estos imponen para la posterior recolección de opiniones asociadas a cada tópico.

Una vez recuperados los documentos desde los sitios de noticias, se procede a utilizar el modelo LDA para recuperar qué tópicos se están tratando en ellos. Este modelo permite, dada una colección de documentos $\{d_i\}_{i=1...\mathbb{N}}$, obtener un conjunto de tópicos t asociados a documentos, los cuales están descrito por la probabilidad $P(topic = t|document = d)$ de que un documento d pertenezca al tópico t y además, para cada tupla (w, t) la probabilidad $P(topic = t|word = w)$ de que una palabra w describa al tópico t. Así, es posible obtener los tópicos que se tratan a lo largo del tiempo en los feeds que se están minando y las palabras que los describen para luego utilizar esta información con el fin de recuperar documentos opinados desde las redes sociales.

Para cada periodo t_i, se toman todos los documentos de los dos periodos anteriores t_{i-1}, t_{i-2} y se entrena un nuevo modelo LDA con estos. Luego, para los documentos del periodo t se realiza inferencia con el modelo LDA sobre estos para descubrir el modelo de tópicos subyacente en estos.

Una vez que se tengan los documentos de los periodos t_i, t_{i-1}, t_{i-2}, es posible enlazar dos tópicos T y T', con vectores de probabilidades de palabras $\vec{w}_T$ y $\vec{w}_{T'}$ través de una función de distancia de tópicos que se define como sigue:

$$d(T, T') = \sum_{w \in \vec{w}_T} \sum_{\vec{w}_{T'}} w_i - w_j \tag{7}$$

Y luego, dado toda dupla T y T' de tópicos, se enlazan sí y sólo si el resultado la función $d(T, T')$ está bajo un umbral ϕ que se define a la hora de comenzar el análisis.

3.1.2. Minado de Opiniones

Una vez que se han extraído los tópicos a partir de los documentos presentes en sitios de noticias, se procede a extraer las opiniones sobre cada uno de ellos en las redes sociales a través del uso de modelos de minado de opiniones. Con esto es posible es posible obtener un puntaje de opinión para cada tópico a lo largo del tiempo a partir de una colección de documentos. La metodología a utilizar es la siguiente:

Para definir qué documentos serán recuperados desde las redes sociales, se tiene el algoritmo 3.2, que dado un tópico T obtiene todos los n-gramas de largo n que caracterizan a ese tópico en particular en el periodo t. El parámetro N consiste en la cantidad de palabras que deben ser utilizadas para la obtención de los unigramas que describen al tópico, además, el método T.`words`(t, N) obtiene las N palabras más relevantes del tópico T en el periodo t.

Algoritmo 3.2: Método `generateQueries`

Data: T, t, n, N
Result: $\{query_i\}_{i \in \mathbb{N}}$

```
1 queries := [];
2 words = T.words(t, N);
3 forall p ∈ permutaciones(words, n) do
4     queries.append(p);
5 return queries;
```

Luego, para cada tópico T, se obtienen todas las queries que le correspondan, y se obtienen documentos opiniones en las redes sociales que se determinen utilizando sus APIs. En el caso particular de este experimento, sólo se trabajará con la red social de microblogging Twitter. Para cada documento, se procede a obtener su polaridad de la siguiente manera:

Algoritmo 3.3: Clasificación de documentos opinados

Data: $\{d_i\}_{i=1...\mathbb{N}}$
Result: $\{\vec{d_i}\}_{i\in\mathbb{N}}$
documents := [];
for $i \leftarrow 1$ **to** $\|\{d_i\}_{i=1...\mathbb{N}}\|$ **do**
    $\vec{d_i} \leftarrow$ `polaridad`(d_i);
    documents.append$(\vec{d_i})$;
return documents;

3.1.3. Visualización de Tendencias

Una vez obtenidas las noticias y las opiniones relacionadas a los tópicos en discusión, se procede a generar un gráfico como el presentado en la figura 1 que representa el comportamiento de cada tópico a lo largo del tiempo. En donde las barras corresponden a la cantidad de documentos en los cuales se hace mención del tópico para cada periodo de tiempo, y la opinión de los usuarios de las redes sociales a lo largo del tiempo con respecto a este se representa como una línea.

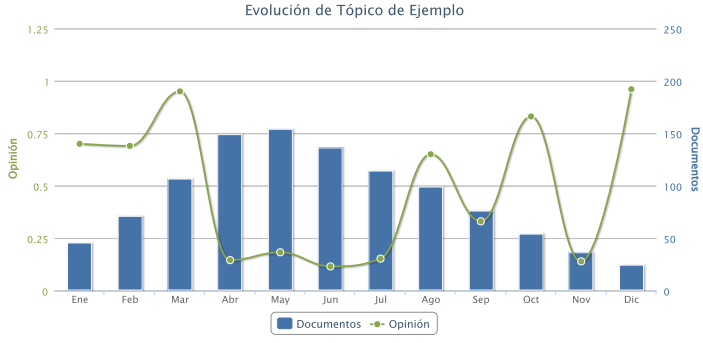

Figura 1: Ejemplo de gráfico por tópico

3.2. Diseño del experimento

Sobre una temática en particular, se visitará periódicamente un conjunto de 20 sitios de noticias que publiquen documentos sobre esta, y se ejecutará la metodología presentada a lo largo de un mes.

3.2.1. El entorno

Los sitios de noticias. En cuanto a los sitios, se requiere satisfacer tres requerimientos: en primer lugar, debe tener una frecuencia de publicación adecuada. Además, la cantidad de documentos por sitio no puede ser excesivo y, por último, estos deben estar en inglés para facilitar el procesamiento de los documentos.

El tema a analizar. El tema a analizar debe ser capaz de generar discusiones en las redes sociales, o al menos muestras de apreciación o desagrado, ya que de manera contraria no será posible realizar la última etapa del proceso de detección de tendencias y por lo tanto, el experimento se verá invalidado.

3.2.2. Captura y transformación de datos

Sitios de noticias. Una vez elegidos los sitios, estos serán visitados periódicamente para recuperar los artículos publicados en ellos. Cada artículo será almacenado con su contenido original, y para su procesamiento se procederá a remover todo el contenido que no sea texto plano (por ejemplo, *tags* de **html**) y todas las stopwords que se encuentren.

Opiniones. Al igual que los artículos recuperados de los sitios de noticias, estos serán almacenados tal como fueron extraídos desde su fuente.

3.3. Soluciones existentes para detección de tendencias

En el ámbito académico, múltiples investigaciones [10, 1, 5] han abordado la detección de tendencias en la web, principalmente en las redes sociales, destacándose entre ellas dos tipos distintos, aquellas que tienen como objetivo detectar de manera temprana aquellos tópicos que serán tendencia en el corto plazo, y las que buscan detectar aquellos tópicos que están siendo tendencia y su presencia va en aumento a lo largo del tiempo.

En aplicaciones comerciales, la plataforma web *NewsWhip*[1] ofrece prestaciones similares a las presentes en la plataforma de detección de tendencias presentada, sin embargo, su enfoque es lograr ser un agregador de noticias con características sociales, como la medición de menciones en las redes sociales de una noticia en particular o el análisis de noticias de una empresa en particular en la web. Además, *NewsWhip* ofrece la herramienta *Spike*, que permite a los generadores de contenido analizar cómo sus noticias se esparcen por la web.

La empresa *Sysomos*[2] se enfoca en monitorear las redes sociales en búsqueda de información relevante para una empresa en particular, sin embargo, no

[1] `http://www.newswhip.com/`
[2] `http://www.sysomos.com/`

hacen uso de la información presente en las noticias y no tienen como objetivo hacer un análisis extenso de las tendencias en la Web, si no monitorear las conversaciones que se están realizando en las redes sociales.

Otra iniciativa que busca detectar tendencias en la Web es Google Trends, la cual toma un enfoque distinto a los ya mencionados al analizar el comportamiento de búsqueda de los usuarios de su motor de búsqueda, sin embargo, no hacen uso de los datos presentes en su red social Google+ para complementar las tendencias obtenidas con información sobre las opiniones de la gente sobre ellas.

4. Aplicación del experimento y análisis de resultados

4.1. Captura de datos

Para recuperar los documentos existentes en los sitios de noticias o blogs que se analizaron, se implementó un *crawler* hecho en Java capaz de parsear y recuperar información desde fuentes *RSS*. Para cada fuente *RSS*, se solicita periódicamente la lista de artículos presente en ella, y en caso de que se encontraran nuevos elementos en relación a la última extracción de documentos se procede a almacenar esta diferencia en la base de datos. En el caso de los documentos opinados recuperados desde las redes sociales, también se desarrolló un *crawler* en Java para recuperar los documentos opinados asociados a un tópico en particular.

4.2. Aplicacion del Modelo de Detección de Tendencias

4.2.1. Entorno

La temática escogida: Los experimentos se desarrollaron con el fin de analizar lo sucedido en la temática de la tecnología y sus ramificaciones, en particular, se enfocó el estudio sobre noticias y opiniones en inglés. Ambas elecciones se realizaron en base a la alta cantidad de información disponible sin importar el periodo en el cual se realizara en el estudio.

Los sitios analizados: Se escogió de manera manual una muestra de 20 blogs o sitios de noticias en inglés que traten la temática de la tecnología. Cada uno de estos debe disponer de su contenido en formato RSS para una más fácil recuperación de sus artículos.

experimento	10	20	30
primero	66 %	58 %	49 %

Tabla 1: *Precision*

experimento	10	20	30
primero	37 %	46 %	59 %

Tabla 2: *Recall*

El periodo de análisis: Para el desarrollo del análisis se analizaron sitios de noticias entre Abril del 2011 y Enero del 2012, analizando los tópicos tratados por ellos en dicho periodo.

4.2.2. Experimentos

Como primer experimento, se aplicó la metodología presentada en el entorno previamente descrito, a partir del cual se procedió a analizar los tópicos extraídos y los gráficos temporales para cada uno de estos, y se determinó si la información presentada en ellos correspondía a lo que se podía observar a partir del análisis de los hechos ocurridos en este periodo. Por otro lado, el segundo experimento consistió en el análisis experto de estos gráficos para ver si dichos tópicos podían ser categorizados como tendencias.

4.3. Resultados Obtenidos

Luego de analizar los sitios de noticias previamente elegidos durante el periodo de análisis con un número variable de tópicos por periodo, y considerando una semana por iteración de la metodología, se encontró que la cantidad de tópicos extraídos por el modelo LDA en cada periodo que ofrecía mejores resultados correspondía a 10 y además, se determinó hacer uso de periodos de 7 días de largo.

4.3.1. *Precision y recall*

En la tabla 1 se muestra la precisión lograda en el primer experimento para las tres cantidades de tópicos por semana que fueron seleccionadas. Se puede observar que a medida que la cantidad de tópicos por periodo aumenta, la *Precision* del algoritmo disminuye, ya que a medida que esta variable aumenta, la granularidad del modelo LDA aumenta, provocando que un tópico descubierto por inspección sea dividido en dos tópicos más pequeños pero altamente relacionados. Este suceso ocurre en todo dominio que se quiera analizar, sin embargo, a medida que el dominio bajo análisis es más amplio, la cantidad

óptima de tópicos por periodo aumenta. Por lo tanto, es necesario ajustar el modelo dependiendo del dominio bajo análisis.

En la tabla 2, se observa que el *Recall* aumenta a medida que la cantidad de tópicos por periodo aumenta. Esto se debe a que si bien hay una mayor fragmentación de macrotópicos, se incluyen tópicos pequeños independientes que son absorbidos por ellos cuando la cantidad de tópicos por periodo disminuye.

En el caso del segundo experimento, se observó que un 63 % de los tópicos observados tuvieron un comportamiento similar al esperado por los expertos consultados, lo que indica, que a pesar de que la herramienta es propuesta como un apoyo a la detección de tendencias, esta realiza un buen trabajo en modelar el comportamiento de los tópicos a lo largo del tiempo.

5. Conclusiones

En este trabajo de investigación se demostró que es posible hacer uso de una herramienta de detección de tendencias basada en datos presentes en la web, para mejorar la calidad de la información provista por medios tradicionales de detección de tendencias como lo son las encuestas de opinión.

Para lograr este resultado se realizó un amplio estudio de cuáles de los datos originados en la web pueden complementar la información presente en los medios tradicionales, junto con los modelos matemáticos que se usan para describir tópicos en colecciones de documentos y la manera en que los usuarios de la web expresan sus opiniones en las redes sociales.

Si bien esta metodología es un complemento para los medios tradicionales, una de sus limitaciones es que la demografía de los usuarios de Internet, y aquellas personas accesibles a través de encuestas no siempre coinciden, por lo que si se desean realizar estudios enfocados en ciertos sectores de la población es posible que esta metodología no logre aportar suficiente valor. Por otro lado, al realizar el análisis de los datos de manera periódica, no es posible dar alerta temprana de sucesos que ocurren en el día a día. Por ello, los resultados entregados por esta herramienta deben ser considerados como un apoyo a decisiones de negocio enfocadas en un mercado en particular y también como un complemento a metodologías tradicionales de detección de tendencias.

Como trabajo futuro, se plantea considerar nuevas técnicas de minado de opiniones que se especialicen en documentos obtenidos desde sitios de microblogging y además características de estos como la ironía y los acrónimos de expresiones populares. Por otro lado, se plantea modificar el modelo de tópicos usado para que sea capaz de detectar reapariciones de tópicos después de

un tiempo prolongado. Finalmente, se propone la evaluación del impacto de implementar un sistema de alerta temprana.

Agradecimientos: Este trabajo fue parcialmente financiado por el Proyecto FONDEF project D10I- 1198: WHALE: Web Hypermedia Analysis Latent Environment y por el Instituto Milenio Sistemas Complejos de Ingeniería (ICM: P-05-004-F, CONICYT: FBO16).

Referencias

[1] F. Alvanaki, M. Sebastian, K. Ramamritham, and G. Weikum. Enblogue: emergent topic detection in web 2.0 streams. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, SIGMOD '11, pages 1271–1274, Athens, Greece, 2011. ACM.

[2] Nikolay Archak, Anindya Ghose, and Panagiotis G. Ipeirotis. Show me the money!: deriving the pricing power of product features by mining consumer reviews. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '07, pages 56–65, San Jose, California, USA, 2007. ACM.

[3] David M. Blei and John D. Lafferty. Correlated topic models. In *NIPS*, 2005.

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, March 2003.

[5] Irena Pletikosa Cvijikj and Florian Michahelles. Monitoring trends on facebook. In *Proceedings of the 2011 IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing*, DASC '11, pages 895–902, Sydney, Australia, 2011. IEEE Computer Society.

[6] T Damer. *Attacking faulty reasoning: a practical guide to fallacy-free arguments*. Wadsworth/Cengage Laerning, Australia Belmont, CA, 2009.

[7] Shay David and Trevor John Pinch. Six degrees of reputation: The use and abuse of online review and recommendation systems. *First Monday*, July 2006. Special Issue on Commercial Applications of the Internet.

[8] Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. Twitter sentiment analysis : The good the bad and the omg ! *Artificial Intelligence*, 70(2):538–541, 2011.

[9] Bing Liu. Sentiment analysis and subjectivity. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, Florida, USA, 2010. ISBN 978-1420085921.

[10] Michael Mathioudakis and Nick Koudas. Twittermonitor: trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, SIGMOD '10, pages 1155–1158, Indianapolis, Indiana, USA, 2010. ACM.

[11] T. Mullen and N. Collier. Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, volume 4 of *EMNLP '04*, pages 412–418. ACL, 2004.

[12] V. Ng, S. Dasgupta, and SM Arifin. Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 611–618. Association for Computational Linguistics, 2006.

[13] B. O'Connor, R. Balasubramanyan, B.R. Routledge, and N.A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, ICWSM '10, pages 122–129. AAAI Press, 2010.

[14] Bruno Ohana and Brendan Tierney. Sentiment classification of reviews using sentiwordnet. *Discovery*, page 13, 2009.

[15] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, January 2008.

[16] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[17] V. Sehgal and C. Song. Sops: stock prediction using web sentiment. In *Data Mining Workshops, 2007. ICDM Workshops 2007. Seventh IEEE International Conference on*, ICDMW '07, pages 21–26, Omaha, Nebraska, USA, 2007. IEEE Computer Society.

[18] Edison M. Taylor, Cristián Rodriguez, Juan D. Velásquez, Goldina Ghosh, and Soumya Banerjee. Web opinion mining and sentiment analysis. In

Juan D. Velásquez, Vasile Palade, and Lakhmi C. Jain, editors, *Advanced Techniques in Web Intelligence-2*, pages 105–126. Springer, 2012.

[19] Peter D. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424, Philadelphia, Pennsylvania, 2002. Association for Computational Linguistics.

[20] Athena Vakali, Maria Giatsoglou, and Stefanos Antaris. Social networking trends and dynamics detection via a cloud-based framework design. In *Proceedings of the 21st international conference companion on World Wide Web*, WWW '12 Companion, pages 1213–1220, New York, NY, USA, 2012. ACM.

[21] Juan D. Velásquez. Web site keywords: A methodology for improving gradually the web site text content. *Intelligent Data Analysis*, 16(2):327–348, 2012.