Universidad de San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Estudios de Postgrado
Maestría en Tecnologías de la Información y la Comunicación

# OPTIMIZACIÓN DE LA BÚSQUEDA DE INTERESES PERSONALES EN TWITTER UTILIZANDO MODELADO DE TÓPICOS

**Ing. Pedro Rafael Ruiz Porras**
Asesorado por la Mtra. María Elizabeth Aldana Díaz

Guatemala, septiembre de 2015

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA


FACULTAD DE INGENIERÍA

# OPTIMIZACIÓN DE LA BÚSQUEDA DE INTERESES PERSONALES EN TWITTER UTILIZANDO MODELADO DE TÓPICOS

TRABAJO DE GRADUACIÓN

PRESENTADO A LA JUNTA DIRECTIVA DE LA
FACULTAD DE INGENIERÍA
POR

**ING. PEDRO RAFAEL RUIZ PORRAS**
ASESORADO POR LA MTRA. MARÍA ELIZABETH ALDANA DÍAZ

AL CONFERÍRSELE EL TÍTULO DE

**MAESTRO EN TECNOLOGÍAS DE LA INFORMACIÓN Y LA COMUNICACIÓN**

GUATEMALA, SEPTIEMBRE DE 2015

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA


## NÓMINA DE JUNTA DIRECTIVA

| | |
|---|---|
| DECANO | Ing. Pedro Antonio Aguilar Polanco |
| VOCAL I | Ing. Angel Roberto Sic García |
| VOCAL II | Ing. Pablo Christian de León Rodríguez |
| VOCAL III | Inga. Elvia Miriam Ruballos Samayoa |
| VOCAL IV | Br. Narda Lucía Pacay Barrientos |
| VOCAL V | Br. Walter Rafael Véliz Muñoz |
| SECRETARIA | Inga. Lesbia Magalí Herrera López |

## TRIBUNAL QUE PRACTICÓ EL EXAMEN GENERAL PRIVADO

| | |
|---|---|
| DECANO | Ing. Pedro Antonio Aguilar Polanco |
| EXAMINADORA | Dra. Mayra Virginia Castillo Montes |
| EXAMINADOR | Ing. Marlon Antonio Pérez Türk |
| EXAMINADOR | Ing. Estuardo Enrique Echeverría Nova |
| SECRETARIO | Ing. Hugo Humberto Rivera Pérez |

## HONORABLE TRIBUNAL EXAMINADOR

En cumplimiento con los preceptos que establece la ley de la Universidad de San Carlos de Guatemala, presento a su consideración mi trabajo de graduación titulado:

**OPTIMIZACIÓN DE LA BÚSQUEDA DE INTERESES PERSONALES EN TWITTER UTILIZANDO MODELADO DE TÓPICOS**

Tema que me fuera asignado por la Dirección de la Escuela de Estudios de Postgrado, con fecha septiembre de 2015.

Pedro Rafael Ruiz Porras

USAC
TRICENTENARIA
Universidad de San Carlos de Guatemala

**Escuela de Estudios de Postgrado**
**Facultad de Ingeniería**
**Teléfono 2418-9142 / Ext. 86226**

APT-2015-032

Como Coordinador de la Maestría en Tecnologías de la información y la comunicación y revisor del Trabajo de Tesis titulado **"OPTIMIZACIÓN DE LA BÚSQUEDA DE INTERESES PERSONALES EN TWITTER UTILIZANDO MODELADO DE TÓPICOS"**, presentado por el Ingeniero en Ciencias y Sistemas **Pedro Rafael Ruiz Porras,** apruebo y recomiendo la autorización del mismo.

"Id y Enseñad A Todos"

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
ESCUELA DE POST-GRADO
FACULTAD DE INGENIERIA

MSc. Ing. Marlon Antonio Pérez Türk
Coordinador de Maestría
Escuela de Estudios de Postgrado

Guatemala, Septiembre de 2015.

Cc: archivo
/la

USAC
TRICENTENARIA
Universidad de San Carlos de Guatemala

**Escuela de Estudios de Postgrado**
**Facultad de Ingeniería**
**Teléfono 2418-9142 / Ext. 86226**

APT-2015-032

Como Revisor de la Maestría en Tecnologías de la Información y la Comunicación del Trabajo de Tesis titulado **"OPTIMIZACIÓN DE LA BÚSQUEDA DE INTERESES PERSONALES EN TWITTER UTILIZANDO MODELADO DE TÓPICOS".** Presentado por el Ingeniero en Ciencias y Sistemas **Pedro Rafael Ruiz Porras**, apruebo el presente y recomiendo la autorización del mismo.

*"Id y Enseñad A Todos"*

UNIVERSIDAD DE SAN CARLOS
ESCUELA DE POST-GRADO
FACULTAD DE INGENIERIA
GUATEMALA

MSc. Inga. María Elizabeth Aldana Díaz
Revisor(a)
Escuela de Estudios de Postgrado

Guatemala, Septiembre de 2015.

Cc: archivo
/la

USAC
TRICENTENARIA
Universidad de San Carlos de Guatemala

**Escuela de Estudios de Postgrado**
**Facultad de Ingeniería**
**Teléfono 2418-9142 / Ext. 86226**

APT-2015-032

El Director de la Escuela de Estudios de Postgrado de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer el dictamen y dar el visto bueno del revisor y la aprobación del área de Lingüística del trabajo de graduación titulado **"OPTIMIZACIÓN DE LA BÚSQUEDA DE INTERESES PERSONALES EN TWITTER UTILIZANDO MODELADO DE TÓPICOS"** presentado por el Ingeniero en Ciencias y Sistemas **Pedro Rafael Ruiz Porras,** apruebo el presente y recomiendo la autorización del mismo.

*"Id y Enseñad A Todos"*

UNIVERSIDAD DE SAN CARLOS
ESCUELA DE POST-GRADO
FACULTAD DE INGENIERIA
GUATEMALA

MSc. Ing. Murphy Olympo Paiz Recinos
Director
Escuela de Estudios de Postgrado

Guatemala, Septiembre de 2015.

Cc: archivo
/la

USAC
TRICENTENARIA
Universidad de San Carlos de Guatemala

Escuela de Estudios de Postgrado
Facultad de Ingeniería
Teléfono 2418-9142 / Ext. 86226

Ref. APT-2015-032

El Decano de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer la aprobación por parte del Director de la Escuela de Postgrado, al Trabajo de Tesis de la Maestría en Tecnologías de la información y la Comunicación titulado: **"OPTIMIZACIÓN DE LA BÚSQUEDA DE INTERESES PERSONALES EN TWITTER UTILIZANDO MODELADO DE TÓPICOS"**, presentado por el Ingeniero en Ciencias y Sistemas **Pedro Rafael Ruiz Porras**, procede a la autorización para la impresión del mismo.

IMPRÍMASE.

"Id y Enseñad A Todos"

Ing. Pedro Antonio Aguilar Polanco
DECANO

Guatemala, Septiembre de 2015.

Cc: archivo
/la

## ACTO QUE DEDICO A:

**Mi mamá**

Por ser – como yo le digo – un roble que simboliza los cimientos de mi vida, y la fuerza necesaria para llenarla de orgullo cuando intento ser un hombre de bien, al servicio de los demás. Mamá, sin vos no alcanzaría nada de esto.

**Mi hermana**

Quien con su actitud hacia la vida, su sabiduría de mujer, y su vocación por superarse, me sirve de ejemplo a seguir en todas las actividades que desarrollo.

**Mis sobrinos**

Luis Fernando, Mariana y José Adrián, porque ustedes son mis maestros para saber cómo canalizar tanta alegría cuando yo sea papá. Ustedes han sido catalizadores de mi estrés durante este proceso de desarrollo académico.

**Mi cuñado**

Quien incansablemente trabaja por mantener a mi hermana y a mis sobrinos en el camino del bien.

# AGRADECIMIENTOS A:


**Mis amigos** Por tenerme paciencia y darme palmaditas en la espalda para seguir adelante. Sin ánimos de olvidar a nadie, hay algunos que debo mencionar en específico: gracias Edwing, Roberto, Zepeda y Alonzo. ¡Es tiempo de celebrar!

**Mi asesora** María, sin tu apoyo y tu confianza, este logro no habría sido alcanzado. Gracias por siempre creer en nosotros y por invertir tanto tiempo y esfuerzo.

**Mis compañeros** Michael, Gaby y Estuardo: ¡Lo hemos logrado! A los demás, que por las vicisitudes de la vida postergaron la culminación de esta maestría, les agradezco sus enseñanzas durante el proceso. Gracias Patal, Alice, Allan, Haroldo, Jonatan, Francisco, Héctor, Mónica, Axel, Geson y Débora. Sigan adelante. Fueron parte esencial de esta etapa académica.

**Mis catedráticos** Les garantizo que sus enseñanzas le han dado valor a mi formación como maestro

# ÍNDICE GENERAL

# ÍNDICE DE ILUSTRACIONES

## FIGURAS



## TABLAS

# LISTA DE SÍMBOLOS

| Símbolo | Significado |
|---|---|
| ***D*** | documento |
| ***T*** | tópico o tema |
| ***W*** | palabra |
| **α** | coeficiente de latencia de LDA |

# GLOSARIO

| | |
|---|---|
| **Conteo de frecuencias** | Algoritmo de extracción de información de textos basado en el peso de la frecuencia de una palabra en todos los documentos. |
| ***Corpus*** | En los métodos de extracción de información de texto, corpus es el colectivo de documentos que se analizan. |
| **Facebook** | Red social de comunidades interactivas que permiten actualizaciones de actividades, eventos y vida social de los usuarios. |
| ***Integrated Development Environment (IDE)*** | Herramienta de desarrollo de software para un lenguaje en particular, o una aplicación en específico, como Visual Studio de Microsoft o Eclipse de Oracle. |
| ***Latent Dirichlet Allocation (LDA)*** | Implementación del modelado de tópicos con una variable latente α que indica la distribución de tópicos en el corpus. |
| **Memoria volátil** | Memoria de acceso aleatorio en un computador. En inglés, denominada RAM, en ella se cargan los procesos (programas) que ejecuta el computador. |

| | |
|---|---|
| **Modelado de tópicos** | Algoritmo de extracción de información de textos basada en distribuciones probabilísticas de temas por documento. |
| **Pre procesado de tuits** | Técnicas y métodos de análisis, filtrado y validación del contenido de los tuits de un usuario para análisis posteriores. |
| **Prototipo** | Artefacto funcional que se utiliza para realizar pruebas de concepto o experimentos sobre un modelo que se desea validar. En software, no es un sistema completo y utilizable en la vida real, se utiliza para iniciar la iteración de un sistema. |
| ***Term Frequency Inverse Document Frequency (TF IDF)*** | Implementación del algoritmo de conteo de frecuencias. |
| ***Timeline*** | El conjunto de tuits de un usuario en Twitter. |
| **Tuit** | Actualización de un usuario en Twitter de hasta 140 caracteres y puede hacer mención a otros usuarios de la red, o incluir enlaces a otro tipo de medios. |
| **Twitter** | Red social de microblogging que permite 140 caracteres en las actualizaciones del usuario. |

# RESUMEN

Las redes sociales se han convertido en un medio de comunicación en donde los usuarios ingresan información personal de todo tipo que puede ser utilizada para diversos fines comerciales, profesionales, u organizacionales. Es incuestionable que las redes sociales forman parte de la realidad humana. En ellas las personas comparten eventos importantes, sentimientos, sucesos de impacto mundial, e información de ocio. Estas redes se han convertido en un medio de comunicación importantísimo para cualquier comunidad.

Sin embargo, los esfuerzos para hacer uso de esta información se han encontrado con fenómenos interesantes. Sobresale el hecho que, por ejemplo en Facebook, la gente promueve una personalidad que no es verdadera. Entonces ¿cómo puede obtenerse información que sea útil a partir de las publicaciones en las redes sociales? Existen diversos métodos de extracción de información a partir de documentos de texto que pueden aplicarse a las actualizaciones en las redes sociales como Facebook o Twitter. No obstante, Twitter tiene una ventaja sobre Facebook en cuanto al contexto para realizar análisis de extracción de información. Y es que Twitter no solicita información específica de intereses o actividades a los usuarios, y no permite textos de más de 140 caracteres. Esta es una ventaja porque en primer lugar, los intereses no estarán sesgados por la doble personalidad ya mencionada; y, en segundo lugar, los usuarios de Twitter se han adecuado a escribir sus ideas de manera concisa en pocos caracteres.

Los primeros trabajos para obtener información de Twitter están basados en el conteo de frecuencias, con algoritmos que determinan el peso de una palabra dentro de un tuit en relación con todo su timeline. Inclusive, existen

servicios como TweetCloud que obtiene una nube de palabras utilizando el algoritmo Term Frequency Inverse Document Frequency (TF IDF, por sus siglas en inglés). Pero la desventaja de este proceso es que produce falsos positivos con palabras que sirven de conectores (conjunciones) o descriptores (adjetivos) que proveen poca información. Otra de las desventajas, y quizás la más grande, es que produce palabras individuales y no considera que el lenguaje humano utiliza tópicos o temas para expresarse.

Recientemente, investigadores de la Universidad de Michigan en el 2011, desarrollaron un algoritmo de extracción de información de documentos con base en probabilidades bayesianas y la consideración de que los documentos están compuestos de temas. A este algoritmo se le conoce como modelado de tópicos, y parte de la premisa de que en un conjunto de documentos cada texto estará formado por una distribución probabilística de temas. Los temas, a su vez, estarán formados por un conjunto de palabras distribuídas probabilísticamente. En otras palabras, un texto en un conjunto de documentos estará formado por cierta cantidad de temas y cada tema estará formado por un conjunto de palabras frecuentes. Por ejemplo, en un conjunto de documentos médicos, los temas a tratar serán patologías y tratamientos. Y cada uno de esos temas tendrá palabras frecuentes como *síntomas* o *medicamentos*, distribuídos con probabilidades distintas.

El modelado de tópicos ya ha sido utilizado para la obtención de intereses en las redes sociales, pero estuvieron dirigidos por los intereses explícitamente seleccionados en Facebook que, como ya se mencionó, no es una fuente confiable de intereses personales por la duplicidad de personalidades.

Este trabajo de graduación parte de la implementación del modelado de tópicos para la obtención de intereses personales en Twitter, con el fin de proponer el diseño de un sistema que optimice el proceso, sin tomar en cuenta

otras redes sociales sino que se enfoque única y exclusivamente en las actualizaciones que los usuarios hacen en Twitter.

Para optimizar el proceso de extracción de intereses personales de Twitter, se hizo uso de un prototipo en el que se llevaron a cabo varios experimentos que evaluaron cualitativamente el uso del conteo de frecuencias en comparación con el modelado de tópicos para el proceso. Este prototipo se codificó en el lenguaje Java y en el IDE Eclipse. El modelado de tópicos propuesto por la Universidad de Michigan está desarrollado en una librería hecha en Java llamada Mallet. Adicionalmente, la librería de código abierto Jate, implementa el algoritmo de TF IDF y fue la utilizada para contrastarla con el modelado de tópicos con Mallet. La experimentación también incluyó el pre procesamiento de los tuits, para encontrar si este pre procesamiento impactó o no en los resultados del modelado de tópicos. Además, el código del prototipo fue instrumentado para explicar cuál de los dos algoritmos tenía el mejor desempeño en cuanto a los recursos del sistema de memoria volátil, tiempo de respuesta y uso de procesador.

Los resultados de los experimentos fueron evaluados por los usuarios analizados. Las personas respondieron una encuesta en la que señalaron cuál de los resultados de ambos algoritmos produjeron temas o palabras que estuvieran más cercanos a sus intereses personales. Más del 80 % de los usuarios investigados indicaron que los intereses obtenidos por el modelado de tópicos eran acertados. Es decir, el experimento pudo determinar que, cualitativamente hablando, el modelado de tópicos optimiza el proceso de extracción de intereses personales en Twitter. En cuanto al pre procesamiento de tuits, nuevamente más del 80 % de los usuarios se inclinó por un proceso que filtre y prepare los tuits antes de llevar a cabo la obtención de intereses personales. En relación con el desempeño del sistema, existe un intercambio entre velocidad y calidad. Es decir, mientras que el TF IDF es 100 veces más

rápido que la implementación del modelado de tópicos, los resultados cualitativos indican que este costo de tiempo de respuesta es justificable frente a la precisión de los intereses obtenidos con el modelado de tópicos. El uso de procesador y el uso de memoria volátil son, en términos generales, similares en ambos algoritmos.

Si bien es cierto que se pudo optimizar la búsqueda de intereses personales en Twitter utilizando modelado de tópicos, aún queda un margen de mejora. Además, no todos los usuarios demuestran sus intereses en sus tuits, muchos solo utilizan la red social para mantenerse informados de noticias relevantes en su medio (como el tránsito, por ejemplo), o para emitir opiniones y juicios de la sociedad en la que viven. Es necesario introducir otros procesos para llevar los productos del modelado de tópicos a un nivel más relevante que pueda realmente aplicarse en nichos de mercado específicos. Los resultados obtenidos con el modelado de tópicos siguen siendo solamente un conjunto de palabras. Podría obtenerse oraciones o temas formales, pero la extracción de la información de textos no puede identificar sentimientos u otras imágenes propios del lenguaje como el sarcasmo, la ironía, o la incertidumbre, por ejemplo. Para solventar estas limitantes podría hacerse uso de la computación con base en el ser humano de manera masiva, de tal forma que los resultados del modelado de tópicos fueran evaluados por miles o millones de personas. Será más fácil que un grupo tan grande de personas pueda identificar dichos sentimientos para encontrar intereses en los tuits de la gente.

# PLANTEAMIENTO DEL PROBLEMA

Las redes sociales necesitan entender a sus usuarios para enviarles información diseñada a la medida. Promueven productos con base en los intereses de las personas. Típicamente esto suele ser con fines comerciales, aunque algunas veces se puede aplicar usos sociales y laborales. Facebook lo hace recolectando información directamente del perfil de las personas. Estos perfiles son ingresados personalmente por los usuarios. Además estudian el comportamiento de las personas con base en las fan pages, likes y eventos que los usuarios eligen. Algunos estudios indican que esta fuente no es completamente confiable debido a que las personas muestran una personalidad en Facebook distinta a la real (Sumner 2011). Twitter, en cambio, no maneja perfiles de usuario. Por lo tanto, para conocer qué intereses tienen las personas, consideran las cuentas que los usuarios siguen y las páginas que visitan. Sin embargo, no aplican ningún otro tipo de data mining; así que, Twitter es una buena fuente de información porque los intereses no estarán sesgados. Es decir, los usuarios tienden a exponer sus preferencias en los tuits.

Hay algunas compañías que indican que apoyan la viabilidad de encontrar los intereses de las personas en los tuits. Pero estas típicamente se fundamentan en la frecuencia de palabras en el timeline (Wiu 2008). Esto solo corresponderá a la periodicidad con que una persona utiliza determinadas palabras sin que esto suponga un interés en específico. Por ejemplo, Tweetcloud hace este conteo. Cabe mencionar que este servicio es solo para una cuenta y no barre un conjunto de usuarios con el fin de encontrar aquellos que sean afines a ciertos intereses.

Existen algunos estudios que no solo se limitan al conteo de frecuencias, sino que incluyen, además, una fórmula para darle peso a las palabras en relación con todo el texto. La desventaja es que puede arrojar falsos positivos porque se enfocan en palabras aisladas y no en temas. Además, los posts que las personas hacen en las redes sociales son micro textos que pueden estar relacionados entre sí. No hace mucho sentido analizar únicamente las palabras en los posts, sino el conjunto de temas subyacentes en la red social de una persona (Forss 2014).

Por lo tanto, el presente trabajo de graduación propone la siguiente pregunta principal:

¿Puede mejorarse la búsqueda de intereses personales en Twitter con la utilización del modelado de tópicos?

Adicionalmente, se plantean las siguientes preguntas auxiliares:

- ¿Qué algoritmos de modelación de tópicos se pueden utilizar?
- ¿Qué tipo de contenido se analizará en twitter?
- ¿Qué técnicas de pre procesamiento de tuits se aplicarán para obtener mejores resultados?
- ¿Qué algoritmo de extracción de temas es más eficiente para determinar los intereses personales en Twitter?

# OBJETIVOS

**General:**

Diseñar un algoritmo que optimice la búsqueda de intereses personales en Twitter con la utilización de modelado de tópicos.

**Específicos:**

1. Identificar el algoritmo de búsqueda de temas que sea más eficiente para encontrar intereses personales en Twitter.
2. Seleccionar el contenido de Twitter con las características relevantes para el análisis de modelado de tópicos.
3. Encontrar las técnicas de pre procesamiento de tuits que optimicen la búsqueda de intereses personales.
4. Evaluar la eficiencia en la utilización de recursos de sistema entre el modelado de tópicos versus el conteo de palabras frecuentes para encontrar intereses personales en Twitter.

# RESUMEN DEL MARCO METODOLÓGICO

Esta fue una investigación experimental y cuantitativa, con el apoyo de la investigación cualitativa. Se realizó un experimento con base en un prototipo de software que utiliza implementaciones de los dos algoritmos evaluados: modelado de tópicos y conteo de frecuencias.

La investigación bibliográfica fue útil para comprender y entender el funcionamiento y la utilización de los algoritmos de extracción de información de documentos de texto. Estos algoritmos fueron utilizados para la experimentación con un prototipo de software.

El prototipo incluyó varios experimentos para alcanzar los objetivos planteados, específicamente:

- Comparación del modelado de tópicos versus el conteo de frecuencias para la obtención de intereses personales en Twitter.
- Extracción de intereses personales de Twitter con modelado de tópicos seleccionando contenido específico de los timelines de los usuarios, y comparando los resultados haciendo el mismo procedimiento pero con todo el contenido disponible.
- Obtención de intereses personales de Twitter con modelado de tópicos pre procesando los tuits de las personas para remover palabras de parada, y comparando los resultados haciendo el mismo procedimiento pero sin procesar los tuits.
- Instrumentación del código de prototipo para obtener indicadores de recursos de sistema y entender el desempeño de los algoritmos de modelado de tópicos y conteo de frecuencias. Se evaluaron los recursos de memoria volátil, tiempo de respuesta y uso del procesador.

Los resultados de estos experimentos fueron evaluados por los usuarios analizados, utilizando una encuesta en la que debieron señalar cuál de los experimentos produjo mejores resultados. Este fue el análisis cualitativo de la investigación con el fin de encontrar la optimización del procceso.

Finalmente, se hizo el diseño de un sistema que obtenga los intereses personales de los usuarios de Twitter, en donde se introdujeron los componentes que optimizan este proceso, específicamente, el componente de modelado de tópicos, así como el procesamiento de los tuits.

**Tipo de estudio**

El trabajo aborda principalmente la investigación experimental y cuantitativa para llevar a cabo los objetivos planteados, además de utilizar la investigación cualitativa para evaluar y discutir los resultados obtenidos.

Se trata de una investigación experimental porque se elaboró un prototipo que utiliza los algoritmos de búsqueda de intereses en los tuits de las personas. El primer algoritmo a utilizado es el de LDA (Latent Dirichlet Allocation, por sus siglas en inglés). Este modela las probabilidades de que una palabra en un documento esté relacionada con un tema cualquiera. Aquellos temas con palabras más frecuentes determinan un tópico como tal. El modelo se hace con base en una serie iteraciones de reemplazo de palabras entre todos los documentos.

El otro algoritmo utilizado es el de TF IDF (Term frequency inverse document frequency). Este algoritmo es comúnmente utilizado en aplicaciones web para indexar las páginas con base en su contenido. Obtiene la frecuencia de una palabra en un texto y le otorga un peso en relación con todos los textos de un grupo de documentos.

El experimento consistió en utilizar el prototipo para obtener los posibles temas de interés de los usuarios de Twitter que hayan aceptado formar parte del experimento. Se obtuvieron cinco palabras con el algoritmo TF IDF y cinco temas con el LDA. Los usuarios respondieron una encuesta para evaluar cuál de los dos algoritmos produjo los resultados más congruentes. Además, otros experimentos consistieron en obtener temas de interés de Twitter con el algoritmo LDA, pero modificando los tuits de los usuarios, de tal forma que pudo identificarse el impacto de pre procesar los textos y limitar el contenido a analizar. La encuesta incluyó preguntas en relación con estas técnicas, y se examinó si su uso en el proceso de obtención de intereses personales es relevante o no. Finalmente, el prototipo fue útil para recabar métricas de desempeño de sistema. Específicamente, se midió el uso de procesador y de memoria volátil, con el objetivo de reconocer cuál de los dos algoritmos es el más eficiente.

Se hizo uso de la investigación cuantitativa para examinar los resultados del prototipo. Se describió cuál de los dos algoritmos es más eficiente en cuanto al uso de recursos de sistema, así como la diferencia en la selección de las técnicas de preparación de los tuits.

Finalmente, la investigación cualitativa fue útil para comparar las evaluaciones que los usuarios hicieron de los intereses encontrados. Y, por último, para determinar qué técnicas de preparación del texto de entrada son las adecuadas para el estudio.

**Variables e indicadores**

Tabla I. **Variables e indicadores**

| Variables | Definición | Sub variables | Indicadores | Dimensiones |
|---|---|---|---|---|
| Modelación de tópicos por medio de LDA | Modelo de probabilidades que describe la ocurrencia y proporción de una palabra en relación con un tema dentro de un conjunto de textos (conocidos como corpus). | 1. Corpus (Conjunto de textos de entrada)<br>2. Temas encontrados<br>3. Iteraciones<br>4.<br>5. Coeficiente de latencia | 1. Cantidad textos de entrada<br>2. Cantidad temas encontrados<br>3. Número de iteraciones<br>4. Palabras de parada | Cuantitativas, enteros, naturales y reales (según sea el indicador). |
| Algoritmo TF IDF | Técnica de frecuencia de palabras dentro de un texto en un conjunto de documentos. Contabiliza la | 1. Corpus (conjunto de textos de entrada)<br>2. Palabras frecuentes<br>3. Peso de | 1. Cantidad de textos de entrada<br>2. Cantidad de palabras encontradas | Cuantitativas, enteros, naturales y reales (según sea el indicador). |

| Variables | Definición conceptual | Sub variables | Indicadores | Dimensiones |
|---|---|---|---|---|
| | | palabras en relación con el corpus | 3. Conjunto de palabras de parada | |
| Contenido de los tuits | El contenido compartido en Twitter pueden ser tuits de creación propia, tuits de otras personas (re tuit), hipervínculos a sitios web y fotografías, o hashtags. | 1. Tuit<br>2. Re tuit<br>3. Hipervínculo<br>4. Fotografía<br>5. Hashtags | Tipo de contenido | Cualitativa |

Continúa Tabla I

| **Variables** | **Definición conceptual** | **Sub variables** | **Indicadores** | **Dimensiones** |
|---|---|---|---|---|
| Técnicas de preparación de textos de entrada | Las palabras de entrada de los algoritmos vendrán de los tuits. Estos deben ser preparados para omitir palabras que no proporcionan información alguna como conjunciones, conectores, o sufijos. Adicionalmente, se eliminan espacios en blanco redundantes y se normaliza el formato de las palabras de entrada. | 1. Separación por caracteres<br>2. Normalización de la capitalización<br>3. Uso de palabras de parada<br>Idioma | Tipo de preparación | Cualitativa |

Continúa Tabla I

| Variables | Definición conceptual | Sub variables | Indicadores | Dimensiones |
|---|---|---|---|---|
| Validez de los intereses encontrados | Evaluación que los usuarios realizarán sobre los resultados del prototipo. Los usuarios indicarán si los intereses encontrados con base en su timeline son acertados. | 1. Calidad de los intereses identificados<br>2. Certeza de los intereses identificados | 1. Rango de valoración de los intereses encontrados<br>2. Porcentaje de certeza | Cualitativa (para el rango) Cuantitativa, real, positivo (para la certeza). |
| Desempeño de los algoritmos | Métricas que demuestren la capacidad de procesamiento y el uso de recursos de sistema de los algoritmos utilizados en el experimento | 1. Uso de recursos de CPU<br>2. Uso de memoria volátil | 1. Velocidad de respuesta<br>2. Porcentaje de uso de CPU<br>3. Cantidad de uso de memoria | Cuantitativa, real, positivo. |

Fuente: elaboración propia.

**Fases**

- Investigación documental: durante esta fase se recopilaron fuentes que no solo explican la metodología para optimizar el proceso de obtención de intereses personales en Twitter, sino que también propician una base conceptual de qué algoritmos se utilizaron, las razones por las que se emplearon, y las ventajas que proveen para optimizar el proceso.
    - Métodos de data mining en redes sociales: se investigaron los procedimientos para obtener información de las redes sociales.
    - Aplicaciones de data mining en redes sociales: descripción de la aplicación de extraer información de las redes sociales y la problemática que presentan.
    - Algoritmos de búsqueda de temas: esta es una de las partes fundamentales ya que explica los dos algoritmos a utilizar en la investigación: el conteo de frecuencias y el modelado de tópicos. Una de las finalidades es poder contrastar ambos métodos para luego describir cómo se optimiza el proceso de obtención de intereses personales en Twitter con la utillización del modelado de tópicos.
    - Uso de la Twitter API: una breve descripción de cómo se accede a la información de Twitter a través de la REST API de la red social.
    - Librerías de código abierto de algortimo de búsqueda de temas: se enumeran y explican las librerías que se utilizaron en el trabajo. Estas librerías implementan los algoritmos de TF IDF para el conteo de frecuencias, y el LDA, para el modelado de tópicos. Estos son los componentes del prototipo que demostraron la optimización del proceso estudiado.
- Diseño del experimento: en esta fase se planteó cómo se hizo el prototipo que soporta  la investigación.

- Definición de objetivos: acá se definieron cuáles son los propósitos de cada uno de los experimentos que se realizaron.
- Selección de experimentos
    - Contrastar resultados entre LDA y TF-IDF (experimento principal): este demuestra cómo se optimiza el proceso estudiado con la utilización del modelado de tópicos.
    - Contrastar resultados sin pre procesar tuits: esta fase diseña el experimento que demuestra cuáles son las técnicas de pre procesamiento de tuits que mejoran el proceso de obtención de intereses personales en Twitter.
    - Contrastar resultados sin seleccionar contenido: esta fase describe el experimento que indica cuál es el contenido que mejores resultados provee en la obtención de intereses personales en Twitter.
    - Medición de uso de recursos de sistema: este diseño será útil para identificar la optimización del proceso estudiado en términos de recursos del sistema como el uso de procesador, memoria volátil y tiempo de respuesta.
- Diseño del análisis cualitativo: no es más que el diseño de la evaluación (encuesta) que los usuarios respondieron para validar la precisión y la optimización de la obtención de intereses personales en Twitter al utilizar la modelación de tópicos, seleccionar contenido específico, y aplicar técnicas de pre procesamiento.
- Elección de técnicas de investigación para analizar resultados: con base en las técnicas estudiadas durante el curso de seminario III, se seleccionaron y aplicaron aquellas técnicas que fueron útiles para la discusión de los resultados.

- Desarrollo del prototipo (programación): en esta fase se escribió el código fuente del prototipo que incluye las librerías que implementan los algoritmos de búsqueda de interereses personales en Twitter. La programación también incluyó el desarrollo de los experimentos diseñados en fases anteriores.
- Experimentación: no fue más que poner a funcionar el prototipo para que se ejecutaran todos los experimentos ya explicados y obtener los intereses personales de Twitter.
- Recolección de resultados: en esta fase se tabularon los intereses obtenidos con los experimentos ejecutados en el prototipo. También se crearon las encuestas personalizadas para cada usuario con el que se haya experimentado. Los usuarios examinaron los resultados y los evaluaron utilizando la encuesta ya diseñada. Los resultados de las encuestas también se tabularon. Estos serán los resultados finales con los que se concluye el trabajo de investigación.
- Redacción del informe final: finalmente, se escribió el informe final donde se explican los hallazgos y la optimización que se obtiene en el proceso estudiado al utilizar modelado de tópicos, selección de contenido de Twitter, y el pre procesamiento de tuits.

# INTRODUCCIÓN

Las redes sociales han trascendido más allá de un medio de comunicación. Son una fuente y un recurso de información diario. La utilización de estas redes se ha vuelto casi esencial para la mayoría de la gente. Han transformado la humanidad llevando a la web un sinnúmero de actividades sociales, profesionales y económicas. Es por esta razón que es importante tomar ventaja de la información que generan los usuarios en las redes sociales.

Diversos estudios se han enfocado en obtener perfiles de personalidad con base en los intereses de los usuarios en redes como Facebook (Hughes 2011). Otros trabajos han utilizado diccionarios predefinidos de intereses aplicados a las actualizaciones de la gente en Twitter, pero han tenido limitantes en cuanto a la precisión de sus resultados. Inclusive, está demostrado que las personas tienden a mostrar una personalidad distinta en las redes sociales de la que tienen en la vida real. En el caso de Twitter específicamente, aunque elimina el sesgo anterior porque no solicita perfiles de usuarios, las investigaciones llevadas a cabo se remiten a la obtención de palabras individuales que carecen de contexto y pueden omitir información oculta.

Estas investigaciones se han basado en la extracción de la información, específicamente en las técnicas de análisis de texto. Comúnmente utilizan el conteo de frecuencias que pondera las palabras dentro de un grupo de documentos. El presente trabajo de investigación aborda el mismo problema pero utilizando otro tipo de método de extracción de información conocido como modelado de tópicos. El modelado de tópicos se fundamenta en la distribución de probabilidades que rigen la combinación de temas y de palabras en un grupo de documentos. La ventaja de este método es que no solo pondera palabras,

sino que produce temas que caracterizan al grupo de documentos y, además, genera modelos matemáticos que se adaptan a nuevos textos.

La investigación se llevó a cabo experimentando con la utilización del algoritmo conocido como Latent Dirichlet Allocation (LDA, por sus siglas en inglés). Este es un modelado de tópicos que utiliza probabilidades bayesianas. El LDA fue utilizado en un prototipo desarrollado con herramientas de software libre: Java como lenguaje de programación y Eclipse como entorno de desarrollo. También se utilizó Mallet, una librería para Java que implementa LDA y que fue desarrollada en la University of Massachussets Amherst. Se descargaron tuits de 27 usuarios utilizando la Twitter4J, una encapsulación de la Twitter API en Java. El modelado de tópicos obtuvo los temas de interés de estos usuarios que luego evaluaron la precisión de los resultados. El objetivo ha sido optimizar la extracción de intereses personales en Twitter con la utilización del modelado de tópicos. La obtención de intereses personales de Twitter puede aplicarse para fines sociales, profesionales, y comerciales.

Los capítulos que formarán el documento final son:

El primer capítulo presenta los antecedentes de esta investigación en donde se describen algunos estudios realizados relacionados con la extracción de la información a partir de las redes sociales, sus métodos, fortalezas y debilidades.

El segundo capítulo justifica el presente trabajo, el campo de investigación y la aplicación de la optimización del proceso. Mientras que en el tercer capítulo se delimitan los alcances del producto de esta investigación.

En el capítulo cuatro se describen brevemente varios conceptos que son fundamentales para este trabajo. Aborda los algoritmos de extracción de información que se contrastan en este trabajo. Estos algoritmos son el conteo de frecuencias y el modelado de tópicos. Se describe en un nivel general su

funcionamiento y se discuten las implementaciones utilizadas en la investigación: el Term Frequency Inverse Document Frequency y el Latent Dirichlet Allocation.

En el capítulo cinco se explica la metodología llevada a cabo en este trabajo. Se explica el diseño del prototipo, las librerías utilizadas y el proceso de descarga de tuits. También trata sobre el diseño de los experimentos que sustentan este trabajo de graduación, con detalles de cómo se aplican los algoritmos de extracción de información en cada uno.

El capítulo seis y siete contiene la presentación y discusión de resultados. Se exponen los resultados de las evaluaciones cualitativas y cuantitativas de los experimentos. Además se discuten y analizan las posibles causas de los resultados, la relación que tienen los mismos y los factores a considerar para examinarlos. El diseño del proceso optimizado se detalla en la sección de presentación de resultados.

Finalmente, en las conclusiones y recomendaciones se explica el producto final del estudio y se proponen futuros trabajos con base en las limitantes encontradas durante la experimentación.

# 1. ANTECEDENTES

Diversos estudios se han realizado para describir el uso del contenido de los usuarios en redes sociales para realizar data mining y obtener información útil para fines comerciales, laborales y sociales. En esta sección se describen algunos de estos esfuerzos.

Una de las redes sociales más destacadas es Facebook, ya que con más de 890 millones de usuarios activos al mes, representa una buena fuente de información de las personas. En el 2011 se realizó un estudio que exploró la correlación entre los usuarios de Facebook y los rasgos de personalidad que se encuentran en las cinco *grandes personalidades*: extraversión, apertura al cambio, responsabilidad, afabilidad y neuroticismo (Sumner, 2011). La investigación consistió en analizar con un programa de software los datos de los perfiles de usuario como género, edad, biografía, y citas de texto, por mencionar algunos. Además, se analizaron las palabras utilizadas por las personas en sus estados de Facebook, para encontrar palabras correlacionadas con las cinco grandes personalidades. Finalmente los usuarios debieron contestar un test de personalidad cuyos resultados fueron contrastados con los obtenidos por el software. El estudio confirma que el contenido que los usuarios publican en la red social denota una personalidad distinta a la de la persona en la vida real. Esto puede deberse a que los seres humanos se preocupan por cómo los observan los demás. Es evidente que la obtención de la personalidad en Facebook resultará en datos sesgados y, por lo tanto, no tendrían aplicaciones prácticas.

Otra de las redes sociales más utilizadas en la actualidad es Twitter. A diferencia de Facebook, Twitter no recaba información personal de los usuarios.

Durante el 2014, una investigación utilizó el algoritmo TF IDF (Term Frequency Inverse Document Frequency) para obtener las palabras que estadísticamente tuvieran más peso en el timeline de un usuario de Twitter. Luego, se utilizaron diccionarios de intereses predefinidos y se aplicó un proceso de reconocimiento de nombres de entidades con base en la información del usuario en Facebook (Forss 2014). La combinación del peso de las palabras, su correlación con los diccionarios predefinidos de intereses y la relación con los intereses encontrados en Facebook, demostraron un acierto del 55 % en los intereses obtenidos de Twitter, según la valoración de los mismos usuarios. Sin embargo, este proceso depende de que los usuarios tengan perfiles en ambas redes sociales y, además, como se evidenció en el estudio de (Sumner 2011), los datos de los perfiles de Facebook están sesgados y no reflejan la realidad.

En relación con analizar únicamente la información en Twitter, se encuentra el trabajo de (Wu et al 2008), quienes además de utilizar el algoritmo TF IDF, lo contrastaron con el TextRank, un algoritmo de búsqueda de etiquetas en textos por medio de grafos, basado en el PageRank de Google. Los resultados obtenidos tienen una precisión del 71.6 % con una desviación estándar de 0.6. Pero en vez de producir un conjunto de tópicos, sencillamente retorna las palabras más utilizadas por un usuario.

En otro esfuerzo por recabar información oculta en Twitter se exploró la capacidad de encontrar los rasgos de narcisismo, maquiavelismo, y psicopatía en los tuits de las personas (Sumner, 2012). Estos tres rasgos, conocidos como la tríada oscura de la personalidad, sirven para determinar la sociabilidad de una persona. Este hallazgo podría aplicarse para definir perfiles de trabajo en procesos de reclutamiento. El estudio utilizó la herramienta Waikato Environment for Knowledge Analysis (WEKA, por sus siglás en inglés), para obtener vectores de las palabras más utilizadas en más de 1700 usuarios de Twitter (Hall 2009). Estos vectores fueron analizados por un aprendizaje de

máquina, desarrollado en colaboración masiva (o *crowdsourcing* en inglés), para encontrar los rasgos de la tríada oscura. La relación entre las palabras y las personalidades de la tríada fue evaluada con el coeficiente de correlación de Pearson. Todas las palabras dieron correlaciones positivas no perfectas ($1 < r < 0$). En otras palabras, las palabras utilizadas por las personas en Twitter sí denotan sus rasgos de sociabilidad. No obstante, aunque el modelo es útil para realizar predicciones a partir de un conjunto de usuarios, la mayoría de las veces en las que se aplicaba para individuos producía falsos positivos. Según los mismos autores de la investigación, su estudio es útil para identificar comportamientos grupales, pero no individuales.

Finalmente, en la investigación de (PAK & Paroubek 2010) se desarrolló una aplicación para obtener las opiniones y los sentimientos en Twitter. Esta utiliza la herramienta de código abierto TreeTagger, para etiquetar las palabras utilizadas en los tuits. Una de las ventajas que presenta este método es la capacidad de entrenar la herramienta con base en frases y oraciones simples en un idioma en particular. El análisis de sentimientos se realizó buscando emoticonos y palabras alegóricas a tres categerías de sentimientos: positivos, negativos, y neutrales. Para evaluar los resultados se calculó la prominencia de las palabras y frases obtenidas por la aplicación. Los valores de prominencia fueron cercanos a 1. Esto indica que las frases y palabras obtenidas están altamente relacionadas con los sentimientos analizados. A pesar de esta relación, el estudio no proporciona intereses por usuario, el rango de sentimientos está limitado a tres categorías generales, y limita la sintaxis de los textos en los tuits, ya que con la utilización de frases en la herramienta utilizada se espera que los tuits sean sintácticamente correctos.

# 2. JUSTIFICACIÓN

Este trabajo de graduación corresponde a la línea de investigación de tecnologías de la información y la comunicación para innovar en industria porque propone un método diferente para la búsqueda de intereses personales en Twitter con la utilización del modelado de tópicos. El uso del modelado de tópicos solo ha sido mencionado en trabajos anteriores pero su aplicación no ha sido validada. La innovación reside en optimizar la obtención de intereses personales en Twitter con base en el modelado de tópicos.

Los métodos tradicionales para obtener información oculta de las redes sociales en general presentan sesgos. Si el usuario es quien ingresa un perfil, entonces ocurre una diferencia de personalidades entre la imagen que quieren proveer y la vida real. Si se realiza el típico conteo estadístico de palabras y su peso dentro de un documento, se incurre en desventajas como resultados que son falsos positivos, limitaciones en las posibilidades de encontrar información oculta, y la incapacidad de predecir intereses individuales.

Los algoritmos de modelado de tópicos superan las limitaciones descritas mediante modelos probabilísticos de sucesos poco comunes. Además, se basan en el concepto de bolsa de palabras, en el que el orden de las palabras no es importante, lo cual es particularmente útil al estudiar Twitter, debido a que es una red social que permite ingresar textos que no responden a las normas de gramática ni sintaxis del lenguaje. El modelado de tópicos es aplicable a textos que no han sido analizados, porque se entrena con un conjunto de tuits iniciales. Esto implica una mejora continua en el modelado de temas ya encontrados.

Las aplicaciones en la vida real de la obtención de intereses personales en Twitter pueden ser comerciales, empresariales, y sociales. Los intereses personales comúnmente se explotan para fines comerciales, con la utilización de anuncios dirigidos a personas con intereses específicos. En el ámbito empresarial, conocer los intereses de las personas puede ser útil para encontrar los mejores candidatos a un puesto de trabajo, o para modificar el clima laboral en una organización. Las aplicaciones sociales podrían ser selección de personas para organizaciones de voluntariado, o para movimientos de participación cívica y conciencia social.

Finalmente, los resultados de esta investigación pueden dar lugar a nuevos estudios que busquen la aplicación de reconocimiento natural del lenguaje, una de las ramas de la información y la tecnología útiles en inteligencia artificial.

# 3. ALCANCES

El alcance investigativo de este trabajo es describir la optimización de la búsqueda de intereses personales en Twitter con la utilización del modelado de tópicos. También se hace una comparación entre el modelado de tópicos y el método de conteo de frecuencias por medio de la experimentación con un prototipo.

El alcance técnico se limita a elaborar un algoritmo que descargue el timeline de varios usuarios de Twitter, y demuestre el impacto de pre procesar los tuits, seleccionar contenido específico, y optimice la búsqueda de intereses personales con la utilización del modelado de tópicos. La implementación del modelado de tópicos a utilizar es la Latent Dirichlet Allocation (LDA, por sus siglas en inglés).

El algoritmo es detallado a partir de los resultados de la experimentación con el prototipo que implemente el LDA. No se creó un sistema funcional, solamente un prototipo para validar la optimización del proceso. Los resultados indican la eficiencia del proceso con la utilización de modelado de tópicos.

Este proyecto es un precedente de la utilización del modelado de tópicos en el análisis de Twitter, por lo tanto, puede servir como base para futuras investigaciones que quieran afinar el proceso o generar un sistema funcional.

Todas aquellas instituciones que quieran encontrar segmentos demográficos para fines comerciales, laborales o sociales, se verán beneficiadas con la optimización del proceso.

# 4. MARCO TEÓRICO

## 4.1 Redes sociales

Las redes sociales son aplicaciones web utilizadas como formas de comunicación electrónica en donde los usuarios crean comunidades en línea para compartir información, ideas, mensajes personales, y cualquier otro tipo de contenido como imágenes y vídeo. Pueden entenderse como un fenómeno que ha transformado la comunicación y la interacción humana en todo el mundo. En la actualidad, las redes sociales han impactado muchos aspectos de la comunicación humana y, por lo tanto, han impactado en los negocios (Edosowman 2011). La utilización de las redes sociales se ha vuelto una práctica diaria en la vida de muchos usuarios.

Las redes sociales se han convertido en un mecanismo para mediar interacciones distantes entre las personas, y se han vuelto prominentes en la era de la información, particularmente con la aparición del internet. Todas las redes sociales permiten a sus usuarios seguir la vida de sus amigos, conocidos, y familiares. El número de personas en las redes sociales ha crecido exponencialmente en la última década (Huberman 2008). Las redes sociales poseen millones de usuarios que usan estas plataformas para seguirse entre sí, encontrar expertos en materias específicas e, inclusive, generar transacciones financieras. Es por esta razón que las empresas tratan de explotar estas redes con propósitos de mercadeo, ya que proveen un medio para propagar recomendaciones de bienes y servicios a través de personas que comparten intereses similares.

Entre las más conocidas se pueden citar Facebook, con más de 890 millones de usuarios activos al mes; Twitter, la red de micro blogs que permite compartir estados de no más de 140 caracteres; Instagram, una red social para compartir únicamente fotografías; Pinterest, para seleccionar ideas visuales; y, por último, LinkedIn, una red social enfocada a fomentar las redes entre comunidades de profesionales.

## 4.2 Twitter y Twitter API

Como ya se mencionó, Twitter es una de las redes sociales más importantes en la actualidad. Es una red social masiva orientada a la comunicación rápida (Kumar 2013). Más de 140 millones de usuarios activos publican alrededor de 400 millones de tuits al día. Twitter ha jugado un rol prominente en eventos sociopolíticos como la Primavera Árabe y la ocupación de Wall Street. Twitter también ha sido útil para informar acerca de desastres naturales, como terremotos, por ejemplo.

Es común que los sitios web incluyan actualizaciones de Twitter embebidas para integrarse con mejor fluidez al entorno de sus usuarios. Por esta razón, Twitter permite acceder a su información por medio de la interfaz llamada Twitter API. Se trata de un servicio web basado en instrucciones REST (típicas del protocolo HTTP). Para acceder a sus funciones, se debe autenticar las solicitudes por medio de la llamada al método Oauth. Este devuelve un token que permite indicar a Twitter que las llamadas subsecuentes a la Twitter API son solicitudes seguras y confirmadas, es decir, un *handshake* de alto nivel entre la aplicación cliente y el servidor de Twitter API. Luego de esta transacción, Twitter API provee varios métodos para interactuar con los timelines de los usuarios, siempre y cuando la aplicación cliente que accede a la Twitter API esté registrada como una aplicación válida en la red social. Esto se consigue por medio de configurar una aplicación a partir de una cuenta de

usuario. Esto garantiza que la utilización de la Twitter API es controlada, auditada, y que no se trata de un bot (Kumar 2013).

### 4.3 *Data mining* en redes sociales

Durante los inicios de la Web, el contenido en línea era estático y producido únicamente por las organizaciones que poseían un sitio en internet. A principios de los 2000, el contenido en línea comenzó a ser generado por los usuarios. Uno de los ejemplos más comunes de contenido generado por los usuarios son los blogs. Más adelante, con la aparición de las redes sociales, el contenido en línea fue más fácil de producir y la cantidad del mismo se incrementó exponencialmente. El contenido generado por los usuarios en la actualidad es de diversos tipos, con base en cada red social. Así, por ejemplo, mientras que en Facebook se observan actualizaciones de la vida cotidiana, en LinkedIn se ve información relacionada con profesiones específicas, mientras que en Twitter se observan comentarios, opiniones, y eventos actuales. Cada red social contiene información relacionada con su temática.

Sin embargo, la cantidad del contenido generado por los usuarios es muy alta. La oferta de información en línea se ha sobrecargado de tal forma que encontrar información de calidad para fines comerciales o sociales es una tarea difícil (Agichtein 2008). Se deben buscar mecanismos que permitan obtener información útil, información que pueda ponerse en práctica para aprovechar la comunicación y la interacción humana en línea. Estos mecanismos se les conocen como data mining de redes sociales. Un concepto que trata sobre la aplicación de métodos de extracción de la información basados en textos, probabilidades, y grafos dirigidos (siempre y cuando se entiendan los usuarios como nodos y sus relaciones como aristas) (Cruz 2011).

## 4.4 Extracción de información

Durante los años 90, algunos estudios demostraron que la mayoría de la gente prefería obtener información de otras personas en lugar de sistemas de extracción de información (Information retrieval, o IR, por sus siglas en inglés). Por supuesto, en esa época, la mayoría de la gente utilizaba agentes de viajes para comprar boletos aéreos. Sin embargo, durante la última década, la optimización continua de la extracción de la información ha dirigido a los motores de búsqueda a nuevos estándares de calidad. Las expectativas de los usuarios han cambiado y, por lo tanto, las búsquedas en línea se han transformado en la más frecuente fuente información. Por ejemplo, (Fallows 2004) determinó que el 92 % de los usuarios de internet reconocieron que la red es un buen lugar para obtener información del día a día.

Sin embargo, la extracción de información no comenzó con la Web. Esta evolucionó desde las publicaciones científicas y registros en bibliotecas, hasta expandirse a otras formas de contenido, como publicaciones médicas, periodísticas, y legales. La mayoría de la extracción de información debe lidiar con el acceso a datos no estructurados y de diversos campos profesionales. Recientemente, uno de los principales incentivos para la innovación ha sido la Web, desatando la publicación de contenido a través de decenas de millones de autores. Esta explosión de la información podría dar lugar a confusiones si la información no pudiera ser encontrada, anotada, etiquetada y analizada, para que cada usuario pueda encontrar información rápidamente y que sea útil para sus necesidades (Manning 2008).

Existen diversas formas de extracción de información que incluyen modelos probabilísticos, transformación, preprocesamiento, valuación y agrupamiento de palabras; vectores, clasificaciones y ordenamientos de textos. Cada uno de estos métodos está fudamentado en modelos matemáticos y

estadísticos, con el propósito de no solo describir la información de un conjunto considerable de datos, sino de proporcionar los indicadores con la mejor calidad posible.

## 4.5 *Text mining*

El *text mining* (o minado de texto) es una forma de extracción de la información que surge a partir de la necesidad de analizar el contenido generado por los usuarios. Esta información típicamente es texto que puede carecer de estructura sintáctica y gramatical. El análisis de texto se hace con base en algoritmos cuyo diseño permite la escalabilidad y el dinamismo de aprender los patrones del texto generado en línea (Aggarwal 2012).

Uno de los principales objetivos del *text mining* es encontrar patrones en el texto, para facilitar el análisis y entendimiento de la información, que después pueda ser utilizado para la toma de decisiones. Esto se logra por medio de algoritmos matemáticos y de considerar a los textos como matrices dispersas y de altas dimensiones. Es decir, se consideran las palabras como unidades que sirven de elementos para conjuntos denominados documentos. A su vez, un conjunto de documentos es considerado un corpus. Las palabras que pueden formar un documento pueden ser cientos de miles (Aggarwal 2012), pero en realidad un documento solo contendrá un sub conjunto de esos cientos de miles.

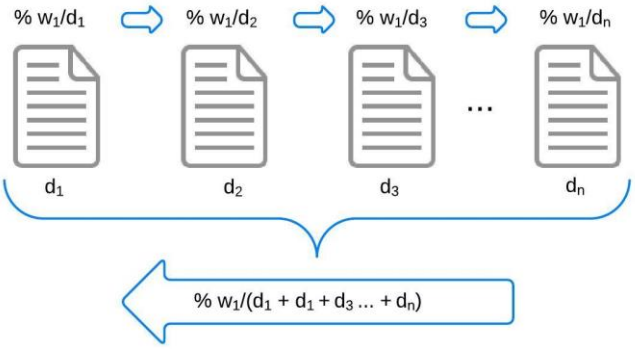
## 4.6 *Term frequency inverse document frequency*

Es un método de extracción de información de textos basado en asignarle un peso a las palabras dentro de un documento y, a su vez, dentro de un corpus. La principal función de darle peso a las palabras es mejorar el proceso de extracción de información. Dos métricas son comúnmente utilizadas

para medir esta función: retención y precisión. La retención se refiere a la capacidad del sistema para extraer las palabras relevantes. Precisión se trata de omitir del análisis aquellas palabras poco frecuentes y, además, poco útiles.

La característica de retención comúnmente se obtiene por medio del factor *term frequency* o TF, porque identifica palabras que son relevantes dentro de un texto. Sin embargo, el factor de TF por sí solo no considera aquellos casos en los que una palabra importante no se encuentra en un solo documento, sino más bien, a lo largo de todo el corpus (Salton 1988). El factor que considera esta distribución es conocido como *Inverse Document Frequency*. La combinación de ambos factores resulta en una ponderación del texto más eficaz y permite obtener términos que pueden definir a un documento.

Figura 1. **Funcionamiento de TF IDF**




Fuente: elaboración propia.

En la Figura 2, se describe en un nivel general el funcionamiento del algoritmo TF IDF. El peso que se le otorga a una palabra dentro de un corpus (conjunto de documentos) está definido por dos conteos de frecuencias. Para cada documento $d_n$ del corpus, se cuenta la frecuencia de la palabra $w_1$ en relación con el total de palabras dentro del documento $d_n$. Luego, la misma

palabra $w_1$ se compara con el total de palabras en el corpus, es decir, la sumatoria de palabras en todos los documentos desde $d_1$ hasta $d_n$. Esta última comparación es la inversión a la que hace referencia el *inverse document frequency* de este modelo. Su aplicación permite mitigar el peso por frecuencia de palabras que aparecen más veces en el corpus, pero que no proveen mayor información. Por ejemplo, las conjunciones y preposiciones son frecuentes pero no tienen un peso importante para el observador. A su vez, incrementa el peso de aquellas palabras que ocurren con menor frecuencia dentro del corpus, pero que están más relacionadas con el tema buscado.

## 4.7 Modelado de tópicos

En años recientes, los investigadores se han visto beneficiados con la digitalización de documentos y textos. Esto produjo una cantidad considerable de información archivada en medios tecnológicos. Con el advernimiento del internet, el acceso a esta documentación se ha hecho factible para la mayoría de las personas. Sin embargo, ocurre la problemática de conocer con precisión en dónde buscar información específica. Se produce el efecto de muchos datos pero poco conocimiento. Es decir, usar esta colección de documentos de manera eficiente requiere de una interacción estructurada y ordenada.

Además, los índices de las ideas contenidas en los documentos y qué documentos tratan de las mismas ideas, no está vigente en la mayoría de las colecciones de documentos. Además la tasa de crecimiento de estas colecciones hace inviable crear estos índices a mano. Para desarrollar las herramientas que permitan explorar estas librerías digitales se necesitan de métodos automáticos que organicen, administren, y entreguen el contenido requerido en un tiempo relativamente rápido.

El modelado de tópicos surge como respuesta a esta problemática. Es un conjunto de algoritmos que utiliza modelos probabilísticos para descubrir la

estructura semántica que existe debajo de todos los documentos que existen en un corpus (Blei 2009). Estos modelos están enfocados en textos, pero pueden ser aplicados en otras ramas como el contenido web, ingeniería genética y las ciencias sociales. Usualmente, el modelado de tópicos aplicado a estas ramas es útil para descubrir agrupamientos entre elementos que no son explícitos. Estos modelos se fundamentan en análisis de textos jerárquicos Bayesianos.

El modelado de tópicos ha sido particularmente útil para encontrar una estructura en documentos que a primera vista no presentan ningún patrón en particular.

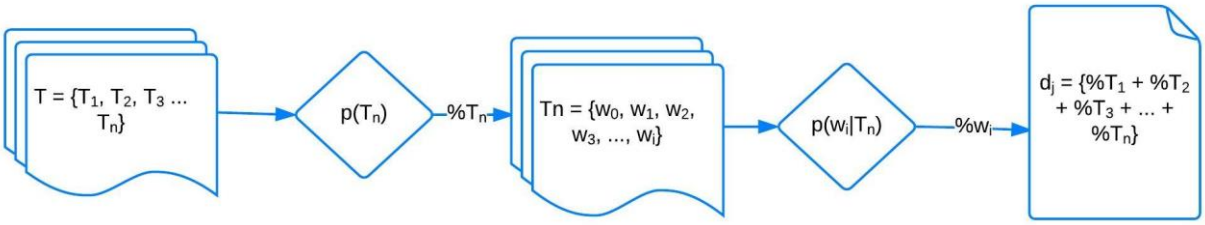
## 4.8 *Latent Dirichlet Allocation*

*Latent Dirichlet Allocation* (LDA, por sus siglas en inglés), es un modelo probabilístico generativo para colecciones discretas de texto, también conocidas como corpus. LDA es un modelo bayesiano de tres niveles jerárquicos, en el que cada elemento en un nivel es modelado como una mezcla finita sobre un conjunto de tópicos subyacente. Cada tópico es, a su vez, modelado como una mezcla infinita sobre un conjunto de probabilidades subyacentes. En el contexto de modelado de textos, las probabilidades de los tópicos proveen una representación explícita de un documento (Blei 2003).

El modelo utiliza probabilidades de Poisson para definir la distribución de temas en el corpus, la distribución de palabras por tema, y la distribución de palabras por tema por documento. Es decir, un tema en específico estará formado por una cantidad discreta de palabras (las palabras no son exclusivas del tema en cuestión, es decir, pueden formar parte de otros temas). Un documento estará formado entonces por una combinación de temas con base en la distribución de probabilidades que tengan los temas sobre el corpus. Las palabras que aparezcan en cada documento estarán en función de la probabilidad de las palabras dentro de un tema en específico.

Es un modelo con un componente latente porque se parte de la suposición de la bolsa de palabras, que se refiere a que las palabras en un documento son intercambiables y, por lo tanto, el orden de las mismas es indiferente para el análisis. Esta propiedad de intercambiabilidad no debe confundirse con aleatoriedad. La distribución de las palabras y su intercambiabilidad estará determinada por una variable. El uso de una variable para la distribución de las palabras (y también de las probabilidades de los tópicos dentro de un documento) le otorga la propiedad de latencia al modelo.

Figura 2. **Funcionamiento de LDA**




Fuente: elaboración propia.

En la Figura 3, se observa que de un conjunto de tópicos desde $T_1$ hasta $T_n$, cada tópico tiene una probabilidad de $p(T_n)$. El porcentaje de distribución de cada $T_n$, determinado por el coeficiente de latencia $\alpha$, dará como resultado la selección de las palabras que conforman el $T_n$. Las mismas palabras $w_0$ hasta $w_i$, tienen una probabilidad de aparecer en un documento condicionada por la probabilidad del tópico $p(T_n)$. Es decir la probabilidad de que la palabra $w_i$ aparezca en el documento $d_j$, es de $p(w_i|T_n)$. Esta probabilidad, basada en la distribución de Poisson, determinará la ocurrencia de una palabra $w_i$ dentro de un $d_j$, para todo $d_j$ que pertenezca a un corpus (conjunto de textos).

## 4.9 Manejo de falsos positivos

En los procesos de extracción de la información es común encontrar datos que pareciera no están relacionados con el contexto del análisis. Es decir, si se toma el caso del modelado de tópicos, y se obtiene un tema formado por palabras que a primera vista no tienen relación directa con los temas del corpus, se dice que se trata de un falso positivo. El algoritmo como tal no puede determinar cuáles resultados son falsos positivos, pero puede indagarse más acerca de su precisión con técnicas de agrupación de variables (AlSumait 2008). Entonces, si se consideran las probabilidades que retorna el modelado de tópicos por cada tema encontrado y se utilizan para, por ejemplo, hacer un análisis de K-medias, se puede inferir qué tan alejado del centro está el falso positivo. Si la distancia es considerable, podría tratarse de un interés emergente o realmente de un tema que no debe considerarse como válido. En el modelado de tópicos implementado en LDA puede ajustarse la variable de latencia alpha, la cual indica la distribución de las iteraciones por realizarse; esta puede ajustarse con más iteraciones que el valor original, y luego se comparan los resultados con los del falso positivo. Existe un punto en el que la cantidad de iteraciones ya no influye en los resultados finales del algoritmo, y donde se puede considerar que el falso positivo ya no debiera suceder. De lo contrario, debe tratarse como un tópico disruptivo.

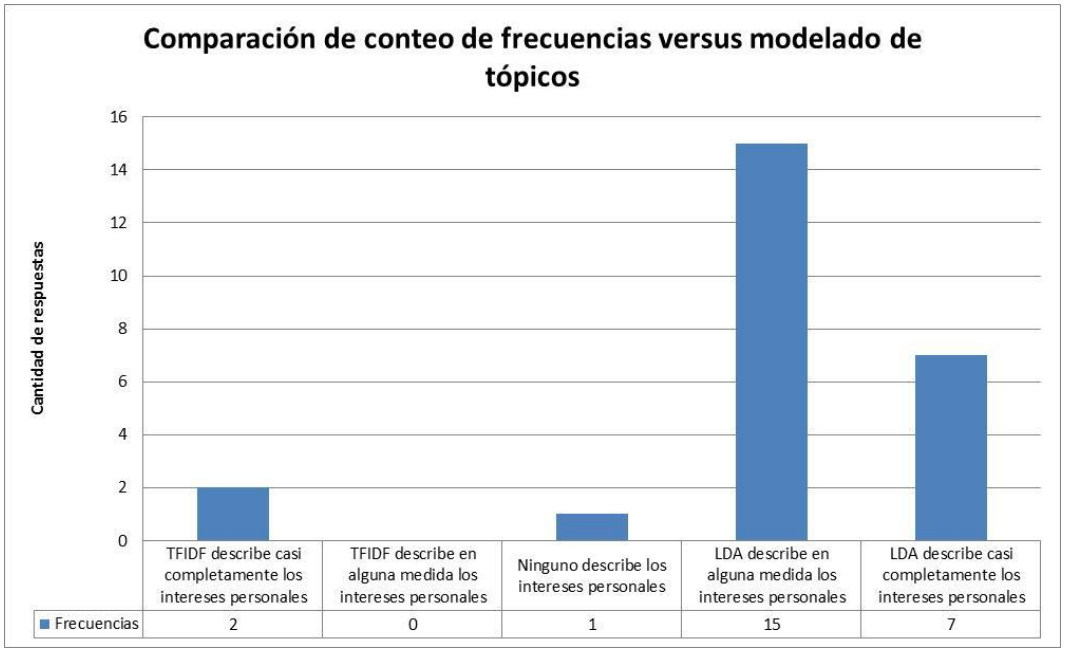
# 5. PRESENTACIÓN DE RESULTADOS

En esta sección se presentan los resultados de los experimentos realizados con el prototipo. Los cuatro experimentos realizados fueron: comparar los resultados de la obtención de intereses personales en Twitter utilizando conteo de frecuencias versus la utilización del modelado de tópicos. Otro de los experimentos fue la comparación de los intereses obtenidos por medio del modelado de tópicos seleccionando contenido original (sin re tuits, enlaces, ni menciones) en contraste con utilizar todo el contenido de los tuits. El siguiente experimento realizado fue contrastar los intereses obtenidos con modelado de tópicos pre procesando los tuits de los usuarios, en comparación con no pre procesar los tuits. Finalmente, se analizaron métricas de desempeño del sistema cuando el prototipo utilizó el modelado de tópicos versus los indicadores del sistema cuando se utilizó el conteo de frecuencias.

## 5.1 Análisis cualitativo

Los intereses obtenidos con el algoritmo de conteo de frecuencias, y los intereses obtenidos con modelado de tópicos, fueron presentados a los usuarios para que evaluaran cuál de los dos describía con mejor certeza sus intereses personales. A continuación los resultados de esta evaluación.

Figura 3. **Resultados cualitativos de LDA vs TF IDF**




Fuente: elaboracción propia.

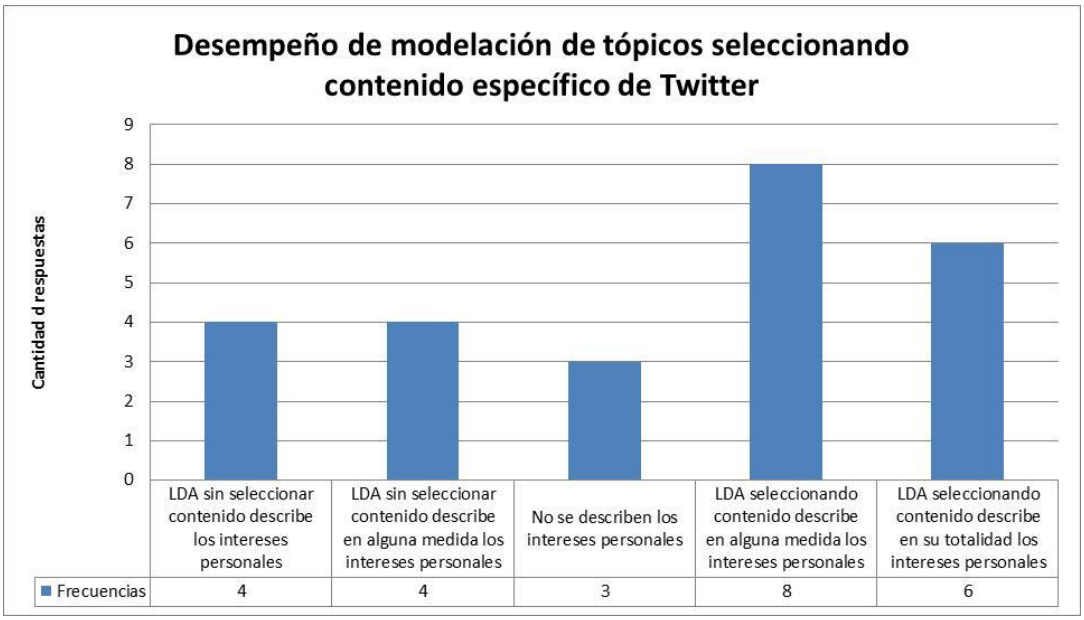
Quince de los veinticinco usuarios indicaron que el modelado de tópicos describe en alguna medida sus intereses personales. Otros 7 usuarios señalaron que los intereses obtenidos con LDA describen completamente sus intereses personales. Es decir, el 88 % de los usuarios analizados se inclinaron por los resultados obtenidos con la utilización del modelado de tópicos como el mejor algoritmo para este proceso.

Dos usuarios indicaron que el algoritmo de conteo de frecuencias describió completamente sus intereses personales.Un solo usuario señaló que ninguno de los algoritmos obtuvo resultados adecuados. Esto significa que el 8 % de los usuarios prefieren el TFIDF como el algoritmo que mejor se adecúa para la obtención de intereses personales.

En relación con seleccionar contenido original de los tuits de las personas para la obtención de intereses personales, los usuarios compararon

los intereses obtenidos con modelado de tópicos sin remover re tuits, enlaces, ni menciones, contra los intereses obtenidos al remover los elementos mencionados. La siguiente gráfica muestra los resultados de esta evaluación.

Figura 4. **Resultados de seleccionar contenido**



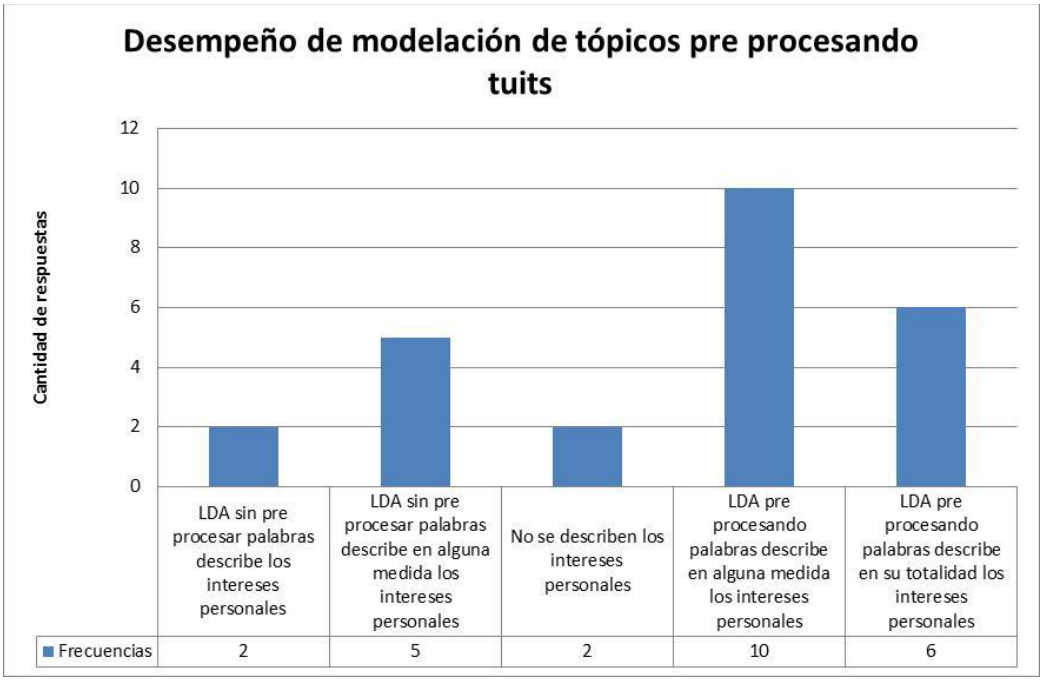

Fuente: elaboración propia.

Tres de los usuarios evaluados indicaron que ninguno de los dos métodos describe sus intereses personales con exactitud. Ocho usuarios señalaron que los mejores resultados se obtuvieron cuando no se seleccionó ningún tipo de contenido en particular de Twitter. Es decir, el 32 % de los usuarios se inclinan por una obtención de intereses sin seleccionar contenido. En contraste, 4 usuarios evaluaron los intereses obtenidos seleccionando contenido como más cercanos a sus intereses personales. Finalmente, un 24 % de los usuarios se inclinaron porque los intereses obtenidos seleccionando contenido describen en su totalidad sus verdaderos intereses personales.

En cuanto a pre procesar tuits para mejorar la obtención de intereses personales, los usuarios analizados evaluaron los intereses obtenidos con el

pre procesado de tuits y los intereses obtenidos sin el pre procesado de tuits. A continuación la gráfica que muestra la evaluación de los usuarios.

Figura 5. **Resultados de pre procesar los tuits**




Fuente: elaboración propia.

Dos de los usuarios indicaron que el modelado de tópicos sin pre procesar tuits describe sus intereses personales en su totalidad. Otros cinco, señalaron que no pre procesar los tuits obtiene intereses relacionados con sus verdaderos intereses personales. Dos usuarios más indicaron que pre procesar los tuits o no es indiferente dado que no se obtienen intereses personales reales. El 64 % de los usuarios, dijeron que pre procesar los tuits para la obtención de intereses personales obtiene resultados que en alguna medida o totalmente describen sus intereses personales.

## 5.2 Desempeño del sistema

A continuación se detallan las mediciones de los recursos de sistema utilizados por las implementaciones de los algoritmos TF IDF (conteo de frecuencias) y LDA (modelado de tópicos). Los recursos examinados fueron memoria volátil (RAM, por sus siglas en inglés), tiempo de respuesta, y uso de procesador (CPU, por sus siglas en inglés).

Tabla II **Uso de memoria volátil (RAM)**

| Iteración | Algoritmo | Valor (MB) | Algoritmo | Valor (MB) |
|---|---|---|---|---|
| **1** | TF IDF | 208 | LDA | 245 |
| **2** | TF IDF | 507 | LDA | 529 |
| **3** | TF IDF | 660 | LDA | 677 |
| **4** | TF IDF | 387 | LDA | 413 |
| **5** | TF IDF | 676 | LDA | 111 |
| **6** | TF IDF | 324 | LDA | 347 |
| **7** | TF IDF | 516 | LDA | 520 |
| **8** | TF IDF | 644 | LDA | 666 |
| **9** | TF IDF | 259 | LDA | 286 |
| **10** | TF IDF | 441 | LDA | 447 |
| **11** | TF IDF | 524 | LDA | 541 |
| **12** | TF IDF | 128 | LDA | 157 |
| **13** | TF IDF | 415 | LDA | 447 |
| **14** | TF IDF | 627 | LDA | 648 |
| **15** | TF IDF | 160 | LDA | 176 |
| **16** | TF IDF | 297 | LDA | 302 |
| **17** | TF IDF | 500 | LDA | 519 |
| **18** | TF IDF | 684 | LDA | 103 |
| **19** | TF IDF | 415 | LDA | 441 |
| **20** | TF IDF | 630 | LDA | 648 |
| **21** | TF IDF | 139 | LDA | 144 |
| **22** | TF IDF | 332 | LDA | 358 |

Continúa Tabla II.

| | | | | |
|---|---|---|---|---|
| **23** | **TF IDF** | **601** | **LDA** | **627** |
| **24** | TF IDF | 285 | LDA | 311 |
| **25** | TF IDF | 544 | LDA | 574 |
| **26** | TF IDF | 235 | LDA | 259 |
| **27** | TF IDF | 542 | LDA | 571 |
| **28** | TF IDF | 209 | LDA | 237 |
| **29** | TF IDF | 314 | LDA | 319 |
| **30** | TF IDF | 570 | LDA | 598 |
| **31** | TF IDF | 227 | LDA | 253 |
| | Promedio | 419.35 | Promedio | 402.39 |
| | Desviación estándar | 176.19 | Desviación estándar | 180.01 |

Fuente: elaboración propia

Luego de realizar el proceso de obtención de intereses personales en 31 cuentas de Twitter, la utilización del modelado de tópicos por medio de LDA utiliza aproximadamente 16.97 MB menos que el conteo de frecuencias TF IDF. La desviación estándar de ambos algoritmos es similar e indica que el uso de memoria volátil en relación con el promedio será de más/menos 180 MB aproximadamente. En otras palabras, el uso de memoria es una métrica dispersa ya que la desviación estándar se aleja en 180 MB del promedio.

Tabla III. **Tiempos de respuesta**

| Iteración | Algoritmo | Valor (ms) | Algoritmo | Valor (ms) |
|---|---|---|---|---|
| **1** | TF IDF | 3,496 | LDA | 30,201 |
| **2** | TF IDF | 870 | LDA | 30,071 |
| **3** | TF IDF | 423 | LDA | 30,201 |
| **4** | TF IDF | 802 | LDA | 30,372 |
| **5** | TF IDF | 785 | LDA | 30,387 |
| **6** | TF IDF | 637 | LDA | 30,361 |

Continúa Tabla III.

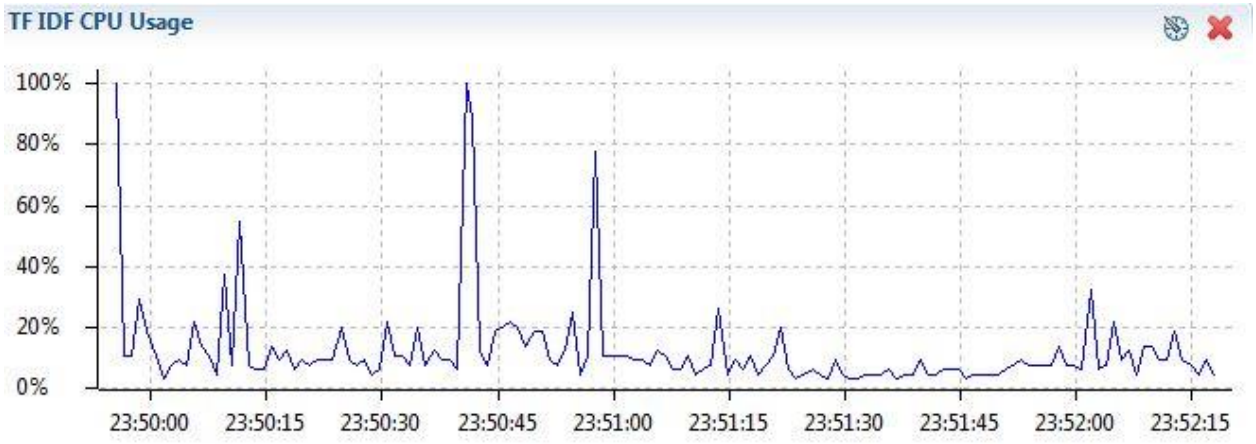
| | | | | |
|---|---|---|---|---|
| **7** | TF IDF | 440 | LDA | 3,013 |
| **8** | TF IDF | 313 | LDA | 30,090 |
| **9** | TF IDF | 521 | LDA | 30,149 |
| **10** | TF IDF | 436 | LDA | 3,021 |
| **11** | TF IDF | 223 | LDA | 3,007 |
| **12** | TF IDF | 549 | LDA | 30,198 |
| **13** | TF IDF | 713 | LDA | 30,256 |
| **14** | TF IDF | 429 | LDA | 30,504 |
| **15** | TF IDF | 360 | LDA | 3,059 |
| **16** | TF IDF | 392 | LDA | 3,075 |
| **17** | TF IDF | 532 | LDA | 30,731 |
| **18** | TF IDF | 502 | LDA | 31,103 |
| **19** | TF IDF | 763 | LDA | 30,697 |
| **20** | TF IDF | 492 | LDA | 30,370 |
| **21** | TF IDF | 243 | LDA | 3,108 |
| **22** | TF IDF | 518 | LDA | 30,481 |
| **23** | TF IDF | 668 | LDA | 30,431 |
| **24** | TF IDF | 762 | LDA | 30,815 |
| **25** | TF IDF | 629 | LDA | 30,755 |
| **26** | TF IDF | 656 | LDA | 30,539 |
| **27** | TF IDF | 760 | LDA | 30,263 |
| **28** | TF IDF | 551 | LDA | 30,526 |
| **29** | TF IDF | 218 | LDA | 3,019 |
| **30** | TF IDF | 860 | LDA | 30,132 |
| **31** | TF IDF | 500 | LDA | 30,104 |
| | Promedio | 646.55 | Promedio | 24,227 |
| | Desviación estándar | 559.82 | Desviación estándar | 11,632 |

Fuente: elaboración propia.

Luego de la obtención de intereses de 31 cuentas de Twitter haciendo uso de los dos algoritmos evaluados (TF IDF para el conteo de frecuencias y LDA para el modelado de tópicos), los tiempos de respuesta de TF IDF son

menores a 1 s (646.55 ms en promedio), mientras que el modelado de tópicos con LDA se toma más de 24 s en promedio. Sin embargo, la desviación estándar de LDA es proporcionalmente más baja que la de TF IDF. Esto indica que los tiempos de respuesta del modelado de tópicos es menos disperso que el de conteo de frecuencias.

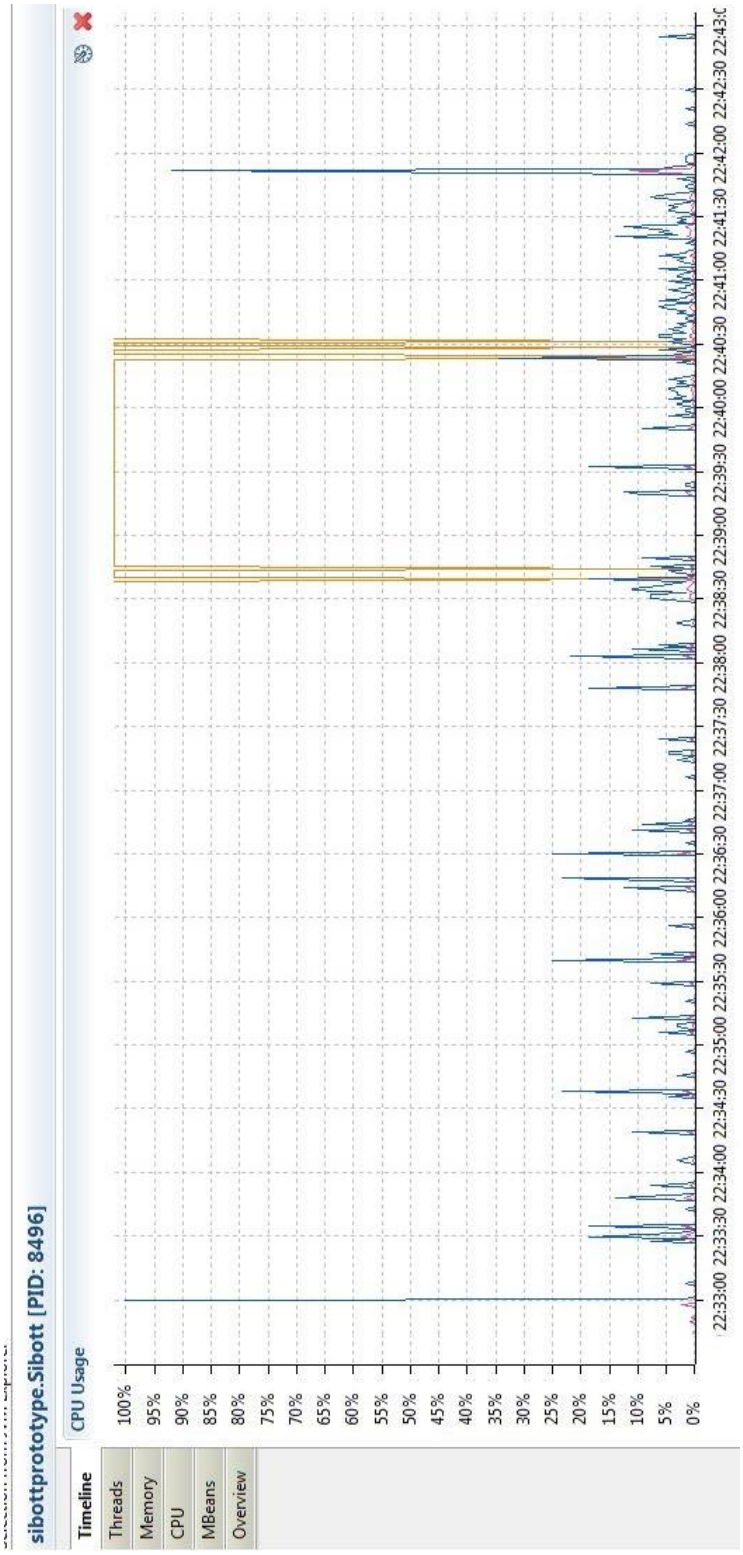
Figura 6. **Uso de CPU de TF IDF**




Fuente: elaboración propia con JVM Monitor en Eclipse.

En la Figura 6: Uso de CPU de TF IDF se muestra la gráfica de utilización del procesador durante el análisis de 30 cuentas de Twitter para la obtención de intereses personales por medio del conteo de frecuencias. En general, la utilización del CPU se mantiene por debajo del 20 %, con algunos picos sobre el 30 y el 40 %, y solo dos picos por arriba de estas mediciones.

Figura 7. **Uso de procesador (CPU) por LDA**



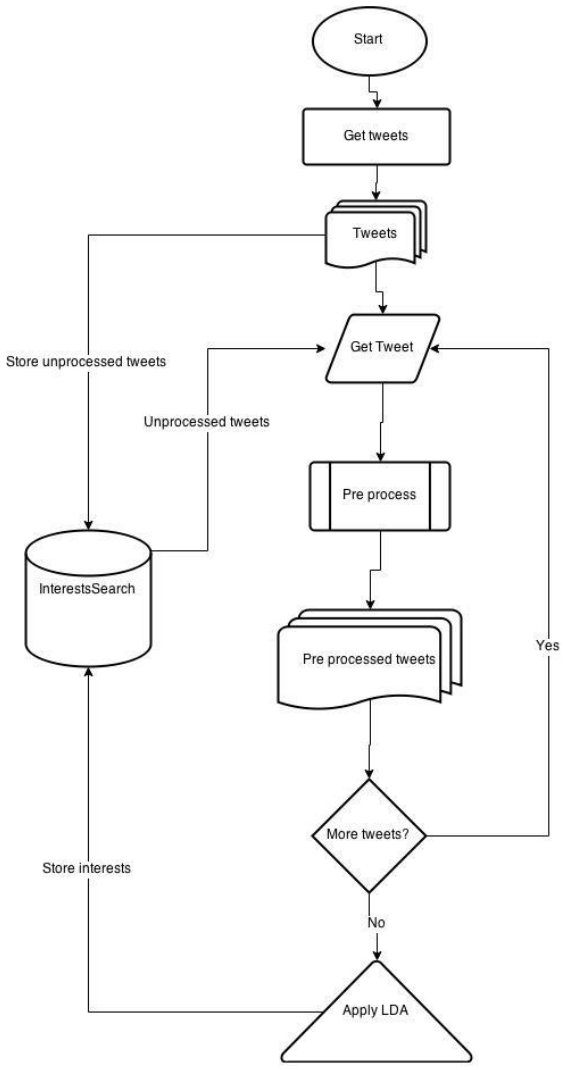

Fuente: elaboración propia con JVM Monitor en Eclipse.

En la Figura 7, se muestra la gráfica del uso del procesador durante la obtención de intereses personales en 30 cuentas de Twitter. En general, el uso del procesador se mantuvo por debajo del 25 %, con algunos picos arriba del 30 %. Inclusive hubo casos en los que se utilizó menos del 5 % del procesador.

## 5.3 Diseño del proceso mejorado

El diseño de un algoritmo que optimice la búsqueda de intereses personales con la utilización del modelado de tópicos se muestra en el siguiente diagrama de flujo.

Figura 8. **Diagrama de flujo del proceso optimizado**




Fuente: elaboración propia.

El proceso de obtención de intereses personales contiene los siguientes pasos:

- Obtención del timeline de cada usuario a evaluar.
- Pre procesar cada uno de los tuits dentro del timeline.
- Si el tuit pasa las validaciones luego del pre procesamiento, se utiliza para el análisis de obtención de intereses personales.
- Cuando todos los tuits del timeline hayan sido pre procesados y validados, se toman como corpus para el análisis de LDA.
- El corpus se analiza con la librería LDA.
- Los tópicos obtenidos se guardan en un medio de almacenamiento para su uso o aplicación.

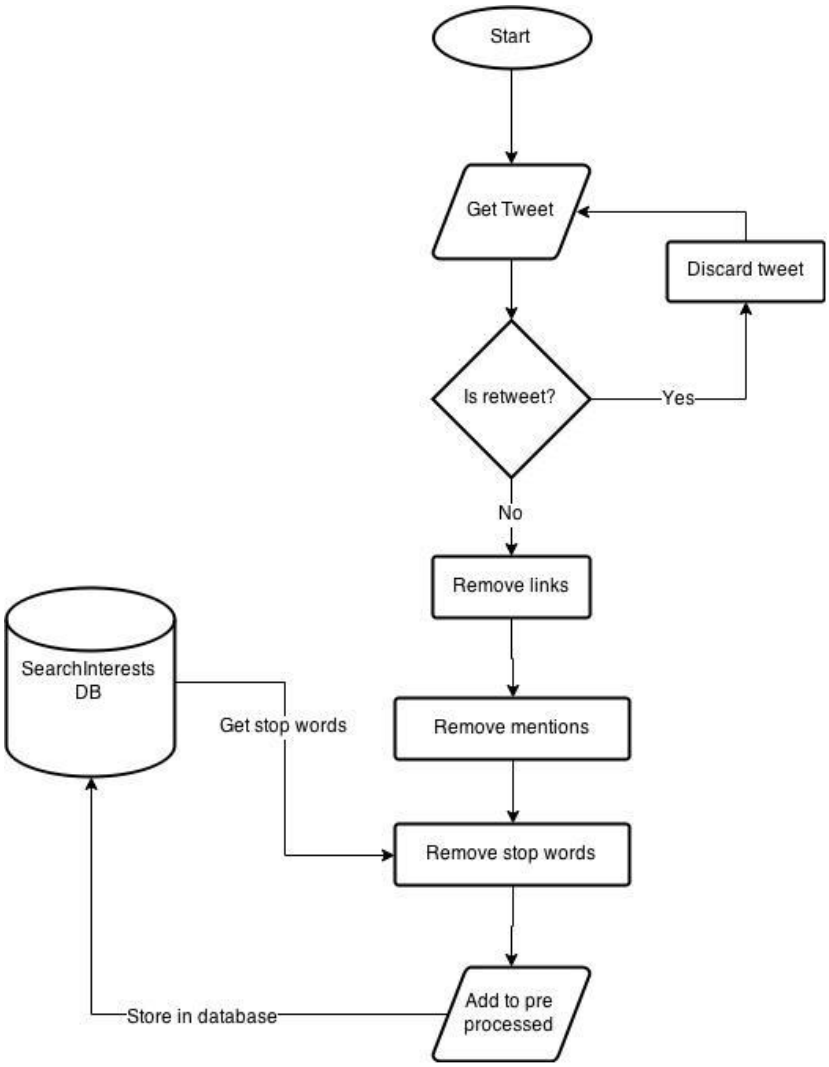
El pre procesamiento y filtrado de tuits contiene los siguientes pasos:

- Se obtiene un tuit del timeline de un usuario.
- Si es un retuit se descarta inmediatamente.
- Si no, se remueven los enlaces. La clase del objeto *Status* (tuit) proveído por Twitter contiene la propiedad *entities*, la cual contiene referencias a los enlaces de sitios e imágenes compartidos por el usuario. Estos enlaces se remueven del tuit.
- Adicionalmente, la clase Status también provee la propiedad *Mentions*, que contiene el listado de usuarios mencionados en el tuit. Al recorrer esta propiedad se pueden remover las menciones del tuit.
- En seguida, se remueven las *stop words* (palabras de parada) del tuit, que pueden estar almacenadas en un archivo de texto o en una base de datos. Existen diversos repositorios en línea de donde se pueden adquirir las palabras de parada en cualquier idioma. Se recomiendan los repositorios de Google o de MySql.

- Una vez se hayan removido los re tuits, enlaces, y palabras de parada, se puede guardar el texto pre procesado en un medio de almacenamiento como una base de datos, en archivos planos, o en una base de datos No Sql.

  El siguiente diagrama de flujo detalla los pasos anteriormente descritos.

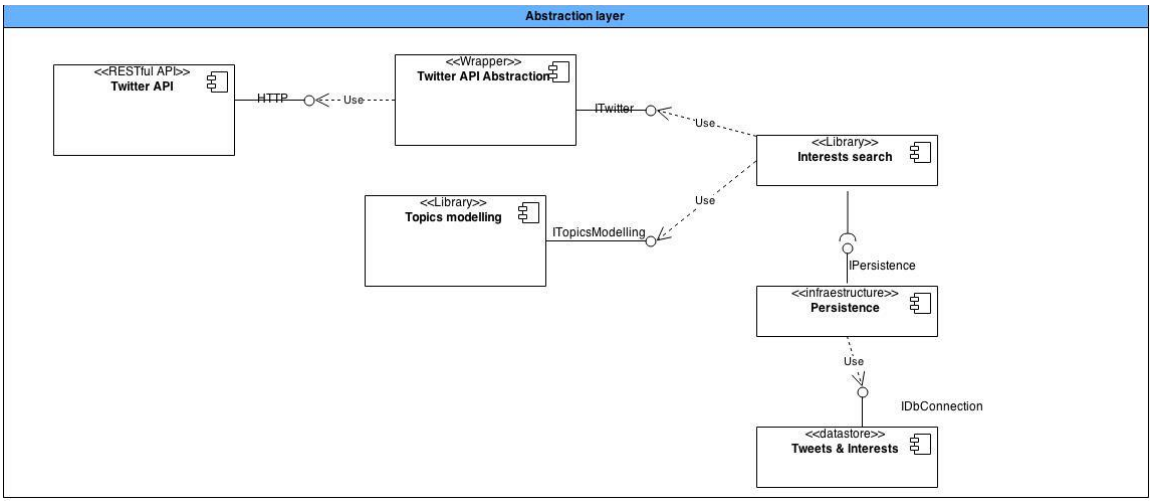
Figura 9. **Diagrama de flujo del procesado de tuits**




Fuente: elaboración propia.

El diagrama de componentes para implementar este algoritmo se describe a continuación.

Figura 10. **Diagrama de componentes del proceso optimizado**




Fuente: elaboración propia.

La Figura 10 muestra el diagrama de componentes que incorpora las librerías para el modelado de tópicos en la obtención de intereses personales en Twitter. La librería de Twitter API es la RESTful API que Twitter provee para acceder a los tuits de los usuarios por medio de aplicaciones individuales y comerciales.

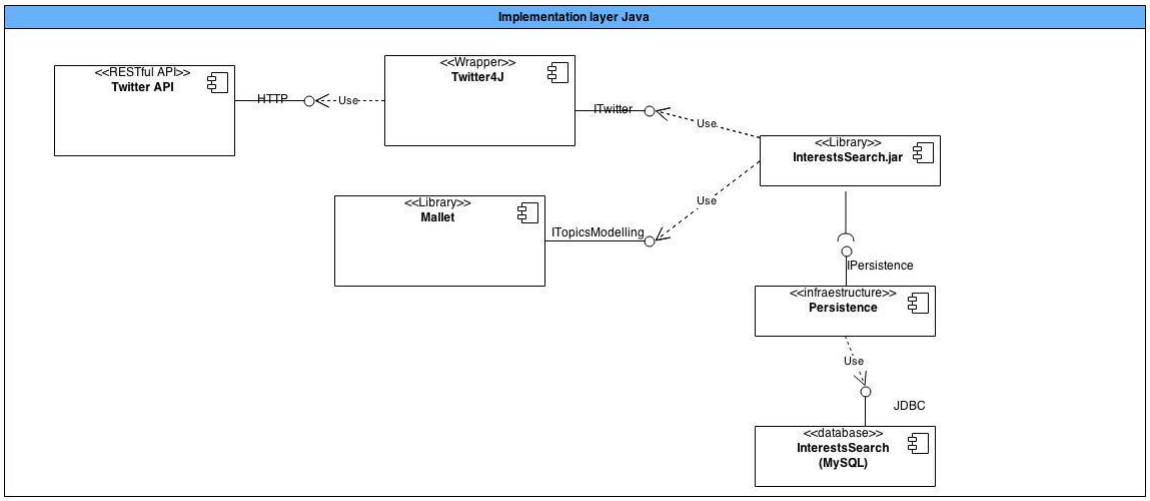
En el mercado existen diversas implementaciones que encapsulan la utilización de la Twitter API para poder implementarse en varias plataformas como .Net de Microsoft, Java de Oracle, o PHP, por ejemplo. El diseño contiene esta librería de encapsulamiento denominada Twitter API Abstraction.

La librería llamada *topics modelling* es la implementación del modelado de tópicos en cualquiera de los componentes existentes del mercado. Esta es la librería que debe utilizarse para aplicar el análisis del modelado de tópicos en el proceso de obtención de intereses personales en los usuarios de Twitter. Esta librería es la que optimiza el proceso.

El componente denominado *Interests search* es la librería principal que hace uso de los componentes descritos. Es la que ejecuta los pasos descritos en el diagrama de flujo de la Figura 8. Orquesta el proceso principal y, además, recupera y almacena los tuits, las palabras de parada y los intereses obtenidos.

Finalmente, los componentes de *Persistence* y *Datastore*, conforman el accceso a los medios de almacenamiento que guardan los tuits, palabras de parada, y los intereses recolectados de los usuarios de Twitter.

Figura 11. **Diagrama de componentes en implementación para Java**




Fuente: elaboración propia.

En la Figura 11 se muestran las implementaciones utilizadas en este trabajo, orientadas al ambiente de desarrollo de Java. La Twitter4J es la librería de código abierto que encapsula la Twitter API. La librería Mallet de la Universidad de Michigan es la librería en Java que implementa el modelado de tópicos con LDA. El InterestSearch.jar sería el sistema formal desarrollado a partir del prototipo utilizado en este trabajo. Como ya se mencionó, es el componente principal que orquesta todos los pasos del proceso optimizado. Para almacenar los resultados, las palabras de parada y los tuits, se utiliza una

base de datos MySql que también es de código abierto, con el conector Java Data Base Connector (JDBC, por sus siglas en inglés).

# 6. DISCUSIÓN DE RESULTADOS

Luego de la obtención de intereses personales y que los usuarios evaluaran los resultados, es evidente que la mayoría valoró los tópicos obtenidos por LDA como de mejor calidad que aquellos recuperados por el TF IDF. Puede decirse entonces que, cualitativamente hablando, el modelado de tópicos mejora considerablemente la calidad de la obtención de intereses personales de Twitter.

No obstante, algunos usuarios indicaron que ninguno de los dos algoritmos brindó resultados que se asemejaran a sus verdaderos intereses personales. Esto puede darse ya que las redes sociales dan lugar a diversos usos que dependen de los usuarios. Es decir, algunas personas utilizan Twitter particularmente para compartir información y noticias útiles, el estado del tránsito vehicular, un evento que ocurre a nivel nacional (las próximas elecciones, por ejemplo), para comunicarse con sus semejantes por medio de las menciones en los tuits, o sencillamente para compartir imágenes u otros tuits relacionados con el ocio. En ninguno de estos casos se verán tuits que denoten verdaderos intereses personales y, por ende, el análisis de los textos en los tuits no proporcionará ninguna información que demuestre afinidades de los usuarios que puedan ser utilizadas para fines comerciales, organizacionales, o profesionales. Particularmente, aquellos usuarios que utilizan los microblogs (como Twitter) para realizar juicios u opiniones, encontrarán que los resultados del análisis de este trabajo proporciona palabras y tópicos que no se relacionan en nada con sus intereses personales. Por lo tanto, no importará qué algoritmo se ejecute, el análisis no recolectará temas útiles.

Por ejemplo, los tópicos modelados con LDA en el siguiente cuadro, provienen de un usuario que indicó que los experimentos no determinaron sus intereses personales. Sin embargo, sus tuits contienen opiniones y juicios de la comunidad en la que vive:

Tabla IV. **Ejemplo de intereses obtenidos con LDA**

| |
|---|
| eso, tan, más, mujeres, ustedes, pone, ternurita, estar, comida, mejor, |
| tráfico, pedazo, tengan, haga, pizza, tanto, perfecto, ese, vida, dicen, |
| gente, mara, expression, alguien, gracias, super, eso, almuerzo, asi, día, |
| hay, personas, amiga, hombre, mejor, más, ejemplo, tipo, diferencia, teoría, |
| alguien, whoa, nada, día, feliz, acuerdo, fijo, forrar, efectos, [redacted]*, |

Fuente: elaboración propia.

*Palabra soez censurada por respeto al lector.

A continuación un extracto del timeline del usuario:

Tabla V. **Extracto de un *timeline***

| |
|---|
| 4,l,Estan peor que los religiosos que venden la Atalaya! Esos solo joden los domingos, estos joden all day every day! |
| 5,l,Me persigue la gente que vende herbalife! Me persiguen por todos lados! En el condominio, en el Facebook, ahora en la oficina! Ya bastaaaa! |
| 6,l,@******_****** estoy 100 % de acuerdo |
| 7,l,RT @******_******: Creo que tiene más sentido que la gente con cierto criterio e inteligencia emocional no busque herir a los demás a propó… |
| 8,l,@ ******_****** si verdad? Es que hay una teoría que dice que si uno es feliz casi que [redacted] arcoiris, no me parece esa teoría |
| 9,l,Será cierto que la gente feliz no trata de herir a otra gente a propósito? |
| 10,l,Y más de alguno está alli no para conocer chav@s o para poner su foto en fb con su camisa, sino para ayudar... digo yo pues. |

Continúa tabla V.

| |
|---|
| 11,l,Y ahora que traen contra un techo para mi país? Yo se que es una moda, pero al menos es una moda que no jode |
| 12,l,#esdegordos http://t.co/8WpOFaHwWj |
| 13,l,Si hablan de ustedes en tercera persona, tienen serios problemas... es en serio. |
| 14,l,Alguien me explica, por favor, que es el "lesbofeminismo" ... qué tiene que ver el feminismo con la orientación sexual pues? |

Fuente: elaboración propia.

En esta muestra del timeline del usuario se puede observar que sus tuits son juicios de la sociedad y contienen varias opiniones al respecto. Adicionalmente, es evidente el tono sarcástico de varias de sus actualizaciones. Una de las desventajas de los procesos de extracción de información es la incapacidad de detectar humor en el texto. Se necesitan otros procesos de inteligencia artificial y reconocimiento del lenguaje natural para alcanzar niveles aceptables de resultados que puedan detectar esos efectos del lenguaje humano.

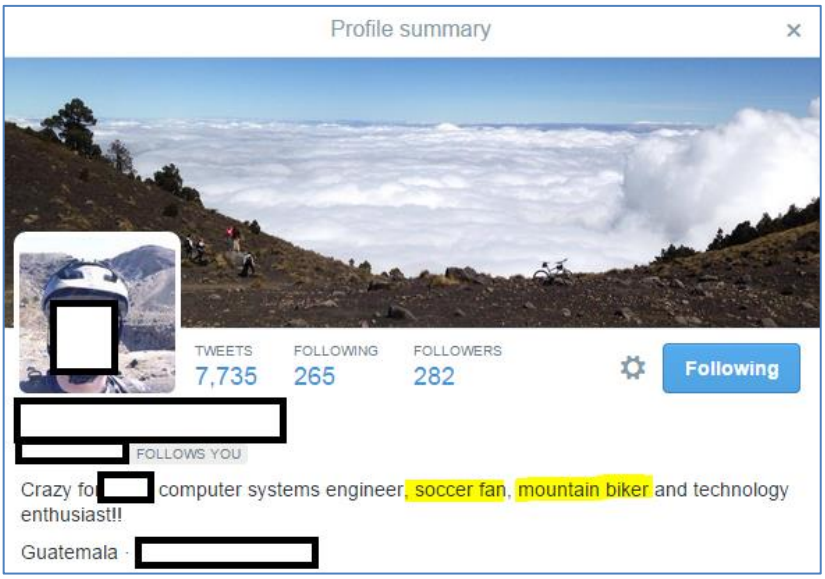
Tabla VI. **Intereses obtenidos con LDA acertados**

| |
|---|
| viscaelbarça, hoy, twitter, día, jajaja, estar, mas, hasta, lloviendo, nunca, |
| rojos, barça, anoeta, [redacted], bienvenido, cara, mountainbike, tocará, psg, pasar, |
| endorphins, golito, traficogt, naranjo, endomondo, walking, finished, casa, [redacted], esa, |
| week, está, how'd, ausopen, fcblive, traficogt, endomondo, walking, finished, salida, |
| partido, twitter, desvelo, balondeoro, oro, abuelos, habría, keoseián, familiares, motivos, |

Fuente: elaboración propia.

En la Tabla VI se muestran los tópicos modelados para otro usuario que señaló el LDA como el mejor algoritmo para obtener sus intereses. Este usuario fue contactado individualmente y explicó que usualmente sus publicaciones en Twitter giran en torno al equipo de fútbol FC Barcelona, la práctica del ciclismo de montaña (mtb, por sus siglas en inglés), y el tránsito en Guatemala. Inclusive y como se ve en su perfil en Twitter (ver Figura 12), sus intereses están bien definidos.

Figura 12. **Perfil de usuario con intereses bien definidos**




Fuente: elaboración propia.

Ambos ejemplos son útiles para explicar que con base en el uso que las personas le den a su cuenta en Twitter se podrán realizar análisis de texto válidos para la obtención de intereses personales con el modelado de tópicos o con cualquier otro algoritmo.

En cuanto a la selección de contenido, el impacto en la obtención de intereses personales nuevamente dependerá del uso que la persona haga con su timeline. Con base en la Figura 4 de la sección de resultados, se observa que un 32 % de los usuarios considera que todo el contenido de sus tuits es valioso para la obtención de intereses personales. Esto puede ocurrir si un usuario re tuitea, o comparte enlaces de sitios, noticias, o tuits que considera de su interés. Puede tratarse de personas que no expresan ideas originales, o no publican actualizaciones de sus actividades personales. Por lo tanto, aunque la mayoría de los usuarios analizados se inclinó por la selección de contenido para mejorar el proceso de obtención de intereses personales, se debe tomar en

cuenta los resultados que puedan encontrarse a partir de tomar en cuenta todo el contenido de los tuits.

En cuanto al desempeño del sistema, se encontraron diferencias entre el algoritmo TF IDF y el LDA para la obtención de intereses personales en Twitter, que están relacionados con el funcionamiento de cada método y con la implementación de las librerías utilizadas en los experimentos.

Para comenzar, ya que el TF IDF de la librería Jate necesita que cada documento dentro del corpus (en este caso, tuit dentro del timeline) exista en un archivo de texto individual, el uso del procesador aparenta ser mayor en comparación con la implementación de LDA en la librería Mallet de la Universidad de Michigan. Esta última librería es flexible en términos de los elementos de entrada (es decir, el timeline de un usuario), ya que permite ingresar los datos por medio de archivos de texto, en conjunto, o uno por uno. Para que la comparación fuese en condiciones iguales, los tuits para LDA también fueron ingresados como archivos de texto individuales. Esta igualdad de condiciones es evidente en el uso de memoria volátil, ambos algoritmos hicieron uso de una cantidad de memoria similar entre los dos, el promedio ronda entre los 400 MB. Sin embargo, LDA se muestra levemente mejor en el uso de la memoria RAM, ya que utiliza más de un hilo de procesamiento para analizar las cuentas de los usuarios. De hecho, la cantidad de hilos a procesar es parametrizable y esto hace que la distribución de la carga de trabajo sea más eficiente en relación con la implementación de TF IDF.

En cuanto al uso del procesador, mientras que los picos del uso de TF IDF fueron mayores que los de LDA, el diseño de la librería Jate que implementa el TF IDF, es un diseño académico y prematuro. Dicho de otra manera, no posee consideraciones de optimizar recursos de sistema y, como se trata de solamente contar frecuencias y pesos dentro del corpus, sencillamente efectúa las operaciones a demanda por cada documento del corpus (tuit del

timeline). Por otro lado, el LDA de Mallet es una implementación más madura con más de 5 años de iteraciones que, por la naturaleza intercambiable del algoritmo de LDA, hace uso de recursos como multi tarea multi hilo. Es por esta razón que Mallet optimiza el uso del procesador con porcentajes de uso más bajos (menores al 15 %), aunque con los picos normales cuando se tratan de timelines extensos. Esto lleva a otra consideración en el desempeño del proceso de obtención de intereses personales en Twitter con modelado de tópicos: la cantidad de tuits de cada persona una vez pre procesados. Existen usuarios asiduos con timelines en el orden de los miles de tuits, y otros usuarios que tienen una cuenta con el único fin de comunicarse por medio de mensajes directos entre usuarios y que, por lo tanto, sus timelines no poseen muchos tuits. La cantidad de tuits es importante cuando se busca examinar el desempeño en términos de uso del procesador de los algoritmos de extracción de información utilizados en esta investigación.

Finalmente, en el tiempo de respuesta TF IDF es más rápido que el modelado de tópicos de la librería Mallet. El análisis de un timeline con alrededor de 250 tuits (después del pre procesamiento) concluía en alrededor de los 30 s cuando se utilizaba el modelado de tópicos. En cuanto que todos los resultados de TF IDF se obtuvieron en el orden de las centésimas de segundo.

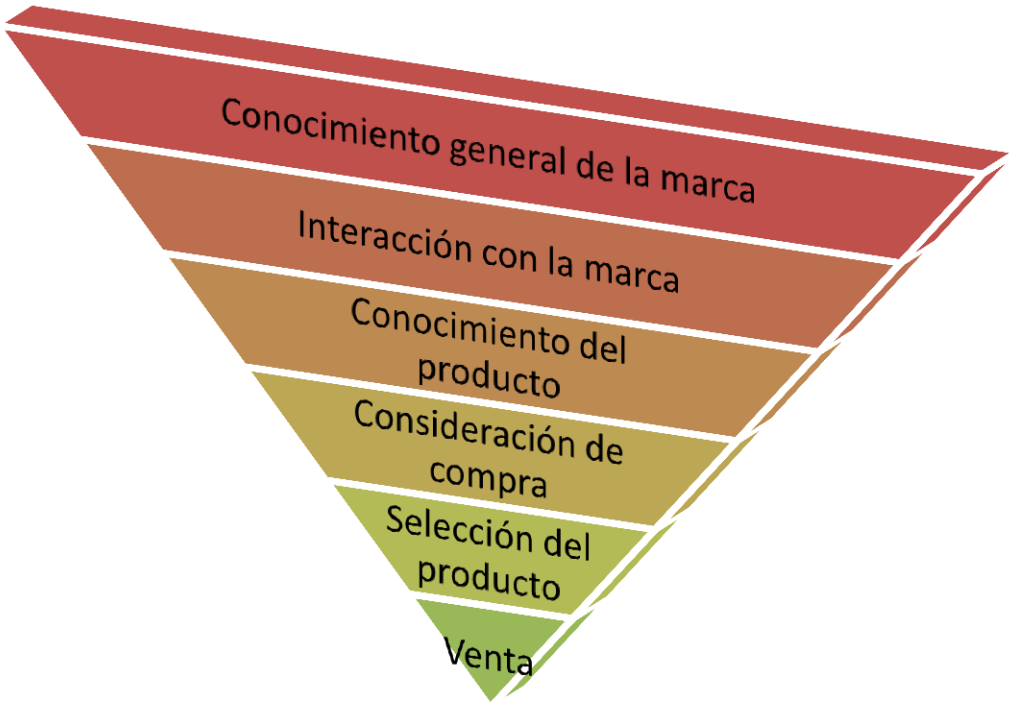
Se observa, entonces, una compensación de recursos entre ambos algoritmos. Es decir, mientras que LDA aparentemente optimiza mejor el uso del procesador con la utilización de técnicas de multi tarea con multi hilos de procesamiento, dado que debe ejecutarse por lo menos 1000 iteraciones sobre cada corpus, los tiempos de respuesta son mucho más largos que aquellos de TF IDF. Por lo tanto, la compensación de recursos se da en la pérdida de velocidad de respuesta con un mejor uso del procesador en el modelado de tópicos. Sin embargo, el valor agregado de la mejora en la calidad de los

resultados justifica el incrementto en la velocidad de respuesta del método propuesto.

## 6.1 Aplicaciones y campo de acción

Estos hallazgos poseen una diversidad de aplicaciones en la vida real. Como se mencionó en la sección de Justificación, las aplicaciones de obtener intereses personales con mayor precisión a partir de una red social como Twitter, permitirían establecer relaciones y agrupaciones de personas para efectos comerciales, en donde se podrían identificar consumidores en el embudo de conversión de ventas a niveles granulares, de tal forma que pueda enviarse anuncios dirigidos con ofertas específicas para demografías especializadas. Puede pensarse en anuncios hacia aquellas personas con intereses específicos en productos particulares, de esta forma, se podría maximizar el retorno de la inversión en términos de costos por anuncios en línea. Se trata de menores costos en anuncios ya que se pagarían anuncios para perfiles de usuario definidos, obtenidos a partir del proceso de obtención de intereses personales con modelado de tópicos, usuarios que potencialmente tendrían más posibilidades de adquirir el producto ofrecido con base en sus intereses.

Figura 13. **Embudo de conversión de compra**




Fuente: elaboración propia.

En un contexto similar, y con base en estudios que indican que los perfiles de trabajo deben alinearse con las expectativas del plan de vida de cada individuo, compañías que busquen reclutar personal para roles definidos, encontrarían una aplicación de alto valor en incluir el análisis de intereres personales en Twitter con modelado de tópicos, porque les permitiría encontrar personas que se apeguen de una manera óptima al perfil de trabajo que se está buscando. Podría, incluso, validarse la aplicación de personas que llegan a través de agencias de bolsas de trabajo, y descartarse de una manera eficiente aquellas personas que definitivamente no apliquen al puesto, o aquellas que llenan un perfil técnico deseado, pero que, con base en los intereses personales obtenidos a partir de las técnicas demostradas en este estudio, podrían encontrarse con un conflicto de intereses que de otra manera no podría

percibirse. Además, la aplicación en el proceso de reclutamiento de personal, con base en los intereses personales, también podría utilizarse para encontrar voluntarios para organizaciones cuya fuerza de trabajo resida en el voluntariado. Quizás, personas que no están decididas a participar en estas actividades, o que tienen la actitud del servicio a los demás pero que sencillamente no conocen a estas agrupaciones, se les pueda extender invitaciones por medio de Twitter, para que se inclinen por finalmente participar en estas actividades que tienen un gran valor humano.

Adicionalmente, para movimientos multitudinarios de índole social, civil, y humano, será útil la obtención de intereses personales en twitter con el modelado de tópicos, porque podrían encontrarse personas que tienen la intención de apoyar movimientos de denuncia política, como el caso de #RenunciaYa durante el año 2015 en Guatemala.

Otro campo de acción es la interacción entre equipos de trabajo en las organizaciones, ya que la mayoría de personas utiliza Twitter para publicar sus sensaciones o actividades durante el día, la aplicación de LDA en la búsqueda de intereses personales, podría encontrar indicadores del clima laboral para que la gerencia tenga mejores indicadores de cómo se percibe el clima organizacional dentro de la empresa. Todo esto a partir del uso que los empleados hagan con la red social. Esta aplicación podría extenderse a realizar tests dirigidos, en los que se les indique a los empleados de una empresa que escriban una o dos páginas relatando su trabajo ideal, y luego utilizar el modelado de tópicos para obtener los intereses comúnes del grupo de trabajo. Esto sería particularmente útil para re alinear la visión corporativa con la visión individual de las personas. Ayudaría a balancear el equilibrio trabajo y vida personal, que se traduciría en una producción mejorada.

## 6.2 Trabajo a futuro

Nuevos estudios podrían surgir a partir de los resultados de esta investigación, ya que, aunque se demostró de forma científica que el algoritmo de modelado de tópicos optimiza la búsqueda de intereses personales en los tuits, aún queda por implementar esta herramienta en ambientes reales, con volúmenes de usuarios considerablemente altos, en los órdenes de los miles, para poder obtener grupos demográficos específicos, basados en los intereses de las personas, y no solo en un cojunto experimental. De esta forma, otro proyecto de graduación podría utilizar los hallazgos de este trabajo e implementar un sistema en línea, con máquinas en la nube, que utilice el diseño plasmado en este documento, que incluya las librerías necesarias (ya sea en código abierto o en código propietario, ya que las librerías existen para diversas plataformas). Esta implementación propuesta podría estar diseñada para incluir elementos de escalabilidad y buscar conjuntos de intereses específicos. Puede pensarse en una base de datos en la que se tenga un catálogo actualizable de intereses objetivo, fundamentados en los tópicos estacionales, para por ejemplo, buscar personas aficionadas a equipos de fútbol durante los finales de temporada y enviarles ofertas de productos relacionados a sus equipos.

Asimismo, y ya que es evidente que solo el proceso de extracción de intereses por sí solo no provee resultados completamente acertados, sino solo aproximaciones (aunque con mejoras en relación con el algoritmo TF IDF), la computación humana podría aplicarse en otro tipo de estudio, siguiendo como ejemplo casos de éxito como la plataforma Duolingo, en la que miles de personas validan las traducciones de textos y sitios web entre diversos idiomas. La idea es que los resultados de LDA sean evaluados y catalogados por miles de personas para poder solventar los obstáculos provocados por las imágenes de sarcasmo, opiniones, y sentimientos encontrados en las publicaciones en las redes sociales. Inclusive, podría agregarse componentes de big data para

almacenar y analizar la información obtenida a partir de la aplicación del modelado de tópicos en la búsqueda de intereses en Twitter. Otra forma de abordar el problema de interpretar el sarcasmo y los sentimientos de las personas podría ser la aplicación de inteligencia artificial a los productos del modelado de tópicos. Por ejemplo, el procesamiento natural del lenguaje (NLP, por sus siglas en inglés), el campo de investigación utilizado en el sistema iOS para la asistente que recibe órdenes por medio de voz, podría aplicarse para mejorar los resultados de la búsqueda de intereses en Twitter con modelado de tópicos. Esta ciencia, al ser aplicada en los productos del LDA, podría ser útil para interpretar las opiniones, los sentimientos, y las figuras del lenguaje que se obtengan de la búsqueda de intereses personales en Twitter.

Además, también podría extenderse el estudio de tal forma que se aplique a las publicaciones de las personas pero en la red social Facebook, y contrastar los intereses obtenidos con aquellos en Twitter; siempre y cuando no se tome en cuenta el perfil del usuario ingresado en Facebook. El objetivo sería determinar si ese sesgo de personalidad influye en los intereses presentados en Facebook.

# CONCLUSIONES

1. El diseño que optimiza la búsqueda de intereses personales en Twitter, debe incluir la comunicación a la red social con la Twitter API, un proceso para filtrar y preparar los tuits, y una librería que implemente el modelado de tópicos.

2. El algoritmo de búsqueda de temas que es más eficiente para encontrar intereses personales en Twitter es el modelado de tópicos.

3. El contenido de Twitter con las características relevantes para el análisis de modelado de tópicos debe ser original, sin enlaces, re tuits, o menciones. Esto es particularmente cierto cuando los usuarios publican actualizaciones originales y no hacen uso del contenido mencionado para expresar sus intereses.

4. Las técnicas de pre procesamiento de tuits que optimizaron la búsqueda de intereses personales incluyen convertir todo el texto a minúsculas y remover las palabras de parada como conectores, prefijos, sufijos, conjunciones, interjecciones, artículos y palabras similares.

5. Se evaluó la eficiencia en la utilización de recursos de sistema entre el modelado de tópicos versus el conteo de palabras frecuentes para encontrar intereses personales en Twitter, y se encontró que el modelado de tópicos es más eficiente en términos de memoria volátil y utilización del procesador al compararlo con el algoritmo de conteo de frecuencias. Sin embargo, el modelado de tópicos posee tiempos de respuesta que

son 100 veces más altos. La calidad de los resultados compensa la velocidad de procesamiento.

# RECOMENDACIONES

1. Utilizar la computación masiva con humanos para interpretar los sentimientos y el sarcasmo de las personas a partir de los conjuntos de palabras obtenidos con modelado de tópicos. Por ejemplo, plataformas como Duolingo o Mechanical Turk de Amazon, hacen uso de la computación masiva con base en humanos para obtener resultados que las máquinas no obtienen.

2. Utilizar código UTF-8 para guardar las palabras de parada con acento. De lo contrario, el dominio de caracteres fallará en encontrar dichos acentos en las palabras. Es importante considerar que las lenguas latinas (español, francés y portugués) hacen uso de estos caracteres especiales.

3. Esta investigación fue realizada con 27 usuarios. Es un número de muestra útil para investigaciones académicas, pero quizás podría agregarse valor si se considerara una muestra más grande, aunque esto supondría una automatización en la evaluación de los resultados.

4. En este trabajo de graduación, las evaluaciones se hicieron por medio de encuestas, elaboradas individualmente por el investigador. Para una muestra considerablemente alta, las encuestas debieran generarse automáticamente.

5. Los intereses personales buscados con el modelado de tópicos no estaban enfocados en ningún producto, servicio, organización u objetivo en específico. Una aplicación en la vida real podría incluir intereses objetivos o dirigidos. La herramienta podría especializarse para cualquiera de los tres ámbitos propuestos: comerciales, profesionales y organizacionales.

# REFERENCIAS BIBLIOGRÁFICAS


1. Aggarwal, C., Zhai, C., (2012), *Mining text data,* New York, USA, Springer.

2. Agichtein, E., Castillo, C., Donato, D., et al. (2008). *Finding high-quality content in social media*. ACM Web Search and Data Mining Conference (WSDM). Conferencia llevada a cabo en California, EEUU.

3. AlSumait, L., Barbara, D., Domeniconi, C. (2008), *On-line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking*, sp, ICDM '08. Eighth IEEE International Conference on Data Mining, Conferencia llevada a cabo en Pisa, Italia.

4. Blei, D., Lafferty, J. (2009). *Topic Models*. Proceedings of the 23rd international conference on Machine learning. Conferencia llevada a cabo en Pennsylvania, EEUU.

5. Blei, D., Ng, A., and Jordan, M. (2003). *Latent Dirichlet allocation*. Journal of Machine Learning Research, 3:993–1022.

6. Cruz, D., et al. (2011). *Entropy Based Community Detection in Augmented Social Networks*. 2011 International Conference on Computational Aspects of Social Networks (CASoN), Conferencia presentada en Salamanca, España.

7. Edosomwan, S., Prakasan, S., Kouame, D. (2011). *The history of Social Media and its Impact on Business*. The Journal of Applied Management and Entrepreneurship, 16(3), sp.

8. Fallows, D. (2004). *The internet and daily life*. Pew Internet & American life project. Recuperado de http://www.pewinternet.org/2004/08/11/the-internet-and-daily-life/ (Abril de 2015).

9. Forss, T., Liu, S., Bjork, K. (2014) *Extracting People's Hobby and Interest Information from Social Media Content*, Terminology and Knowledge Engineering 2014, Jun 2014, Berlin, Germany, pp 9 .

10. Fox, C. (1989). *A stop list for general text*. AT&T Bell Laboratories. ACM SIGIR Forum, 24(1), 19-35.

11. Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., y Witten, I. (2009). *The WEKA Data Mining Software: An Update*. 11(1).

12. Huberman, B., Romero, D., Wu, F. (2008). *Social networks that matter: Twitter under the microscope*. First Monday, 14(1), 1-2.

13. Hughes, D., Rowe, M., Batey, M., Lee, A. (2011). *A tale of two sites: Twitter vs. Facebook and the personality predictors of social media usage*, Manchester, Reino Unido, Elsevier.

14. Manning, C., Raghavan, P., Schutze, H. (2008). *Introduction to information retrieval*. Cambridge, Inglaterra, Cambridge University Press.

15. Kumar, S., Morstatter, F., Liu, H. (2013). *Twitter Data Analytics,* New York, EEUU, Springer.

16. Pak, A., Paroubek, P. (2010). *Twitter as a Corpus for Sentiment Analysis and Opinion Mining,* In Proceedings of the Seventh Conference on International Language Resources and Evaluation, Conferencia impartida en Valleta, Malta.

17. Salton, G. Buckley, C. (1988). *Term-Weighting Approaches in Automatic Text Retrieval*, Manchester, Reino Unido, Elsevier.

18. Steyvers, M., Griffiths, T. (2007) *Probabilistic Topic Models*, en Landauer, T., McNamara, D., Dennis, S., y Kintsch, W. (eds), Latent Semantic Analysis: A Road to Meaning, 430-445, Nueva York, EEUU, Laurence Erlbaum.

19. Sumner, C., Byers, A., Boochever, R., Park, G. (2012). *Predicting Dark Triad Personality Traits from Twitter usage and a linguistic analysis of Tweets*. IEEE 11th International Conference on Machine Learning and Applications ICMLA 2012, Conferencia llevada a cabo en Florida, EEUU.

20. Sumner, C., Byers, A., Shearing, M. (2011). *Determining personality traits & privacy concerns from Facebook activity*, sp, The Online Privacy Foundation, recuperado de https://www.onlineprivacyfoundation.org/research/personality-facebook/ (Abril de 2015)

21. Wu, W., Zhang, B., Ostendorf, M. (2008). *Automatic Generation of Personalized Annotation Tags for Twitter Users*. En Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp 689–692