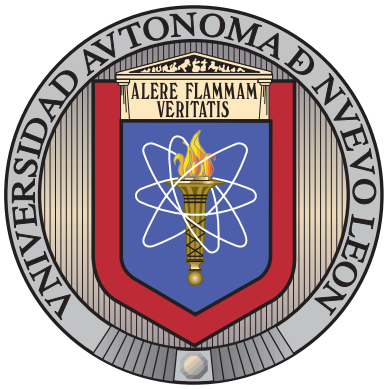
Universidad Autónoma de Nuevo León

Facultad de Ingeniería Mecánica y Eléctrica

División de Estudios de Posgrado



Cuantificación del Interés de un usuario en un tema mediante minería de texto y análisis de sentimiento

por

Fernando Manuel Rodríguez Aldape

en opción al grado de

Maestría en Ingeniería de la Información

con orientación en Inteligencia Artificial

San Nicolás de los Garza, Nuevo León junio 2013

Universidad Autónoma de Nuevo León

Facultad de Ingeniería Mecánica y Eléctrica

División de Estudios de Posgrado

Cuantificación del Interés de un usuario en un tema mediante minería de texto y análisis de sentimiento

por

Fernando Manuel Rodríguez Aldape

en opción al grado de

Maestría en Ingeniería de la Información

con orientación en Inteligencia Artificial

San Nicolás de los Garza, Nuevo León junio 2013

Universidad Autónoma de Nuevo León

Facultad de Ingeniería Mecánica y Eléctrica

División de Estudios de Posgrado

Los miembros del Comité de Tesis recomendamos que la Tesis «Cuantificación del Interés de un usuario en un tema mediante minería de texto y análisis de sentimiento», realizada por el alumno Fernando Manuel Rodríguez Aldape, con número de matrícula 1188110, sea aceptada para su defensa como opción al grado de Maestría Profesionalizante en Ingeniería de la Información con orientación en Inteligencia Artificial.

El Comité de Tesis

Dra. Sara Elena Garza Villarreal
Asesor

Dr. Francisco Torres Guerrero
Revisor

M.C. Aída Lucina González Lara
Revisor

Vo. Bo.

Dr. Moisés Hinojosa Rivera
División de Estudios de Posgrado

San Nicolás de los Garza, Nuevo León, junio 2013

A mi familia, quienes me han brindado su apoyo y amor incondicional.

Índice general

Índice de figuras

Índice de tablas

Agradecimientos

A mi asesora de tesis, la Dra. Sara Elena Garza Villarreal que con su supervisión, paciencia y buena disposición ha sido posible la realización de este trabajo.

A mis revisores de tesis, la M.C. Aída Lucina González Lara y el Dr. Francisco Torres Guerrero quienes dedicaron tiempo y esfuerzo en revisar la tesis y nos apoyaron con su consejo durante este tiempo.

A la Universidad Autónoma de Nuevo León y a la Facultad de Ingeniería Mecánica y Eléctrica por proporcionarnos un lugar para trabajar y aprender.

A las personas que me permitieron utilizar sus cuentas de Twitter para obtener el repositorio de comentarios.

A mi novia Nidia Lizzeth Gómez y mi hermano Jose Adrián Rodríguez, que me compartieron sus experiencias como estudiantes de Maestría— lo cual fue muy enriquecedor.

A mi amigo Rafael Olivares, que dedicó tiempo en la evaluación de los comentarios utilizados en la experimentación al igual que mi asesora de tesis.

A mi madre Ana Cristina Aldape que siempre me apoyó y estuvo al pendiente durante este tiempo.

A toda mi familia que ha estado siempre ahí y me han apoyado incondicionalmente en cada decisión. Sin ellos este objetivo habría sido impensable.

Resumen

Fernando Manuel Rodríguez Aldape.

Candidato para el grado de Maestro en Ingeniería de la Información
con orientación en Inteligencia Artificial.

Universidad Autónoma de Nuevo León.
Facultad de Ingeniería Mecánica y Eléctrica.

Título del estudio:

Cuantificación del Interés de un usuario
en un tema mediante minería de texto y
análisis de sentimiento

Número de páginas: 110.

Objetivos y método de estudio: En la presente tesis se define y propone un método para cuantificar el interés de un usuario en un tema utilizando técnicas de minería de texto y análisis de sentimiento en español. Proponemos utilizar minería de texto para evaluar qué tanto habla un usuario acerca del tema y análisis de sentimiento para saber qué sentimiento utiliza al hablar de éste. Para lograr este último objetivo desarrollamos una herramienta de análisis de sentimiento en español llamada TOM, la cual probó tener una precisión comparable a otras herramientas en el estado del arte. Se experimentó con el método propuesto en un conjunto de 40,186,542 comentarios provenientes de 80,954 usuarios extraídos de un microblog

llamado Twitter. Para cuantificar el interés del usuario propusimos tres estrategias para crear representaciones de los usuarios, las cuales a su vez, pueden variar según el sentimiento de dichos comentarios. Así mismo, propusimos un método simple y automático para representar los temas mediante vectores de palabras y valores TF-IDF. Finalmente mostramos que el Interés, cuantificado mediante nuestro método, tiene una correlación superior a 0.8 en la mayoría de los casos con la probabilidad de que dos usuarios sean contactos en el microblog. Esto concuerda con el principio de homofilia el cuál dice que las personas tienden a contactar con más frecuencia a personas más similares que a personas menos similares [69]. Dicho principio también ha sido estudiado en microblogs [22].

Contribuciones y conlusiones: La contribución principal de la tesis es la definición del Interés del usuario en un tema y el método propuesto para cuantificarlo; el cual tiene en cuenta lo cotidiano o frecuente que es para el usuario el hablar sobre un tema y el sentimiento que utiliza al hacerlo. Entre las contribuciones podemos mencionar las representaciones de los usuarios y los temas, así como la posibilidad de utilizar el sentimiento como un filtro de contenido para crear la representación de los usuarios. Esta idea podría ser utilizada para crear otros modelos que representen los intereses de los usuarios.

Otra contribución importante es la herramienta que desarrollamos para el análisis de sentimiento en español llamada TOM que puede ser utilizada por otros investigadores dada la escasez de este tipo de recursos en nuestro idioma. Una posible aplicación del Interés es la de crear sistemas de recomendación para empresas que deseen promocionar sus productos en un mar de usuarios como lo son los microblogs.

Firma del asesor: _____

Dra. Sara Elena Garza Villarreal

Capítulo 1

Introducción

1.1 El Interés del usuario sobre un tema

Desde los años 90s, el Internet se ha convertido en una opción importante para obtener información y comunicarse. Una de las conclusiones de un estudio realizado por Pew Internet Survey en 2004 [32] afirma que las personas utilizan Internet para tomar decisiones en su vida diaria. Obtuvieron que el 92 % de los usuarios creen que el Internet es un buen lugar para obtener información día con día, el 85 % dijeron que es una buena forma de comunicarse con otros y el 69 % mencionaron que es un buen lugar para entretenerse. En otro estudio realizado por la misma organización en 2010 [50] encontraron que el 58 % de los ciudadanos en Estados Unidos han buscado información acerca de productos en Internet y el 32 % han escrito un comentario sobre un producto. Lo cual es interesante desde un punto de vista comercial.

El Internet está creciendo con mucha rapidez, llegando a incrementarse la cantidad de usuarios en 1,410.8 % del año 2000 al 2012 en América Latina con una penetración en 2012 del 42.9 % de la población; teniendo México una penetración de Internet del 36.5 %[1] [95]. Debido a este crecimiento y la cantidad de contenido que se genera diariamente en Internet, la tarea de buscar información que sea del interés del usuario se convierte en una tarea muy compleja. El usuario estaría frente a miles (o millones) de opciones y le sería imposible evaluarlas todas. Por esta razón han surgido áreas como la *Recuperación de Información* la cual intenta ayudar al usuario

[1] Población en 2012 de 114,975,406 habitantes con 42,000,000 de usuarios de Internet

en la búsqueda de información. Otras áreas como los *Sistemas de Recomendación* buscan hacer recomendaciones al usuario que le puedan resultar de interés con base en un conocimiento de su comportamiento que se utiliza para buscar y evaluar automáticamente otras opciones.

Recientemente han surgido las redes sociales y microblogs, los cuales permiten a los usuarios compartir sus opiniones, sucesos de su vida diaria o información. Son espacios en Internet donde los usuarios pueden comunicarse entre ellos y generar contenidos tales como: como comentarios, conversaciones, fotografías, videos, enlaces a información, etc. Un estudio realizado en Estados Unidos [65] encontró que el 67 % de los usuarios de Internet utilizan alguna red social o microblog; en primer lugar se encuentra Facebook[2] con el 67 % y en segundo lugar Twitter[3] con el 16 %. Entre los usuarios que tienen una edad de 18 a 29 años el uso de alguna red social es del 83 %, siguiendo un 77 % para usuarios de entre 30 y 49 años de edad. Debido al contenido y penetración de las redes sociales y microblogs, éstos se han convertido en fuentes de información muy valiosas desde un punto de vista comercial, ya que tienen un contenido muy rico sobre los intereses de los usuarios que los utilizan [30].

1.2 Justificación y Motivación

Durante décadas se han utilizado técnicas como la encuesta o los grupos de discusión, con las que se pretende conocer las necesidades, opiniones, ideas, emociones y pensamientos de los consumidores [12]. Las empresas invierten dinero y tiempo en publicidad y otros artefactos que les permiten ganar clientes [97]. Requieren hacer investigación de mercados con el fin de conocer a los consumidores para poder afinar sus estrategias de marketing [12]. Sus esfuerzos incluyen encuestas, organización de grupos de discusión, imagen de marca, análisis de posicionamiento de marca, etc. [56]. Estos estudios necesitan tiempo para llevarse a cabo y analizar los resultados, lo cual es una desventaja. El dinero que representa tener que subcontratar un ser-

[2] www.facebook.com

[3] www.twitter.com

vicio de investigación de mercados o hacer la investigación directamente puede ser incosteable [93]. Las redes sociales en Internet que han aparecido recientemente y que han crecido muy rápidamente, representan una nueva forma menos costosa de investigar las necesidades y opiniones de los consumidores [84, 112, 96, 28]. La cantidad de información que se puede obtener de ellas está generando gran interés por parte de investigadores y empresas ya que se obtiene información en tiempo real de las opiniones, deseos, quejas, emociones, pensamientos, relaciones sociales, ubicación geográfica, etc. de los usuarios [8, 109, 39] y es más barato que hacer una encuesta o reunir a un grupo de personas que cumplan con un perfil [116].

Es por eso que en esta tesis deseamos introducir formalmente el concepto de Interés de un usuario en un tema (producto, servicio, marca, etc.) y cuantificarlo númericamente. Una utilidad que se le podría dar a esta métrica es la detección de clientes potenciales en un microblog que estén mas dispuestos a adquirir un producto o servicio.

1.3 Microblogs

Los microblogs son espacios en Internet donde las personas pueden comunicarse e intercambiar opiniones con otros. La principal característica de los microblogs es que se basan en la distribución de mensajes cortos de texto (o comentarios), imágenes con un texto breve, enlaces a video o a información. Los usuarios pueden utilizar estos mensajes para comunicar sucesos de su vida diaria, dar a conocer noticias, vender algún producto o servicio, solicitar algún tipo de ayuda, informarse sobre algún evento, entre muchos otros usos posibles. La mayoría de estos servicios son gratuitos y en los últimos años han tenido un gran crecimiento a nivel global. Como mínimo suelen pedir un correo electrónico para registrarse como usuario y en algunos casos podrían limitar el uso a partir de una cierta edad.

Una vez que se ha registrado, el usuario puede comenzar a buscar contactos con los que puede interactuar. Los mensajes que se publican dentro del microblog

son mostrados al usuario en un área principal, *línea de tiempo* ("timeline") o *canal* donde van fluyendo conforme otros usuarios los publican; generalmente en orden cronológico. La decisión de ver mensajes provenientes de ciertas fuentes depende del usuario. Si éste desea recibir actualizaciones de dicha fuente se convertirá en su contacto. Algunos microblogs llaman *seguidores* ("followers") o *amigos* ("friends") a estos contactos dependiendo de la relación con el otro usuario. Por ejemplo, cuando un usuario desea recibir actualizaciones de otro se convierte en un *seguidor*, pero si ambos desean recibir actualizaciones del otro se diría que son *amigos*. La terminología exacta depende de cada microblog.

La mayoría de los microblogs reciben sus ganancias económicas a partir de la publicidad que es mostrada a sus usuarios y que es pagada por empresas que desean anunciarse. Esta publicidad se enfoca según algunas características del usuario como pueden ser su ubiación geográfica o sus gustos e intereses obtenidos mediante un procesamiento de su información personal publicada en el sitio. También existen microblogs que realizan un cobro para poder registrarse (los cuales podrían no utilizar publicidad dentro del sitio web o hacerlo con poca frecuencia).

Entre algunos microblogs podemos mencionar Plurk[4], Identi.ca[5], Weibo[6], Twitter[7] y App.Net[8] que al momento de la escritura de esta tesis se encuentran vigentes. Otros microblogs como el caso de Jaiku que fue comprado por Google[9] y cancelado en el año 2011[10] han desaparecido.

Plurk: Su lema es: "Regístrate para compartir pequeños mensajes, enlaces, videos y cualquier otra cosa con tus amigos". Se basa en la propagación de mensajes cortos que no tengan más de 140 caracteres. El registro es gratuito pero solicitan un correo electrónico. En sus políticas de privacidad mencionan que desplegarán

[4] www.plurk.com
[5] http://identi.ca
[6] www.weibo.com
[7] www.twitter.com
[8] https://join.app.net
[9] http://www.20minutos.es/noticia/289000/0/google/compra/jaiku/
[10] http://bitelia.com/2011/10/google-cierra-buzz-jaiku

publicidad a sus usuarios basados en la información personal que tienen acerca de ellos.

Identi.ca: Se basa en la propagación de mensajes cortos limitados a 140 caracteres como máximo. Solicitan un correo electrónico para el registro el cual es gratuito. Permite utilizar usuarios de otros microblogs o redes sociales como Twitter o Facebook. Tiene una función para crear grupos de usuarios y compartir mensajes solo con ellos si se desea. Para su creación utilizaron un software de código abierto (Open Source) y aunque mencionan que su giro es comercial, no hemos conseguido encontrar publicidad en él.

Weibo: Es uno de los microblogs más populares en China. Los usuarios pueden escribir comentarios de hasta 140 caracteres. Solicitan un correo electrónico para registrarse, lo cual es gratuito. Su modelo de negocio está basado en publicidad, comercio electrónico y juegos sociales como Farmville[11].

Twitter: Se basa en la propagación de mensajes cortos limitados a 140 caracteres como máximo. Solicitan un correo electrónico para el registro y requieren que el usuario tenga una edad mayor a 13 años. Obtienen ganancias económicas a través de mensajes promocionados los cuales son vistos por los usuarios al momento de realizar alguna buúsqueda. Otra forma de hacer publicidad es a través de temas o tópicos promocionados que aparecen junto a una lista de los tópicos más importantes del momento.

App.Net: Los mensajes están limitados a 256 caracteres de longitud. Su modelo de negocio está basado en la venta del producto (microblog) junto con otros servicios dentro de éste y no en publicidad, por lo que el registro no es gratuito. Requiere que el usuario tenga una edad mayor a 13 años.

[11] http://company.zynga.com/games/farmville

1.4 Definición del problema

Las empresas necesitan de sus clientes para poder sobrevivir y necesitan obtener información sobre ellos para poder conocer el mercado y ser más competitivos. Supongamos que una empresa que lanza un nuevo producto desea buscar posibles compradores en redes sociales o microblogs. Nos haríamos las preguntas, "¿De qué forma podemos saber si un usuario estará interesado en comprar?", "¿Cómo podría obtener una lista de posibles usuarios interesados provenientes de las redes sociales?", "¿Cuáles serían las características de estos usuarios que los hacen diferentes de los demás?", "¿Se podría automatizar el procedimiento de búsqueda para generar un reporte diario?".

En esta tesis abordamos directamente el problema de identificar a las personas en un microblog que con mayor probabilidad adquirirían un producto o servicio, o estarían interesados en conocerlo. Definiremos formalmente el concepto de Interés de un usuario en un tema y presentaremos un procedimiento para medirlo que tendrá en cuenta el contenido del usuario y el sentimiento que se encuentra en dicho contenido. De esta forma podemos cuantificar el agrado o desagrado que tiene un usuario sobre un tema.

1.5 Preguntas de Investigación

Describimos la pregunta principal en la que se concentra el trabajo que se va a realizar en esta tesis, y de ésta se desprende una serie de preguntas que al resolverse contribuirán a su consecución.

- ¿Cómo se puede definir y cuantificar el interés de un usuario en un tema?
  - ¿Es posible hacer análisis de sentimiento en texto en español?
  - ¿Es posible utilizar minería de texto en español?

- ¿Es posible hacer minería de texto en combinación con el análisis de sentimiento de texto en español?
- ¿De qué forma se podrían representar los usuarios?
- ¿Se podría utilizar el sentimiento para crear las representaciones de los usuarios?
- ¿De qué forma se podrían representar los temas?
- ¿De qué forma se podría calcular la similitud del contenido del usuario y un tema?
- ¿De qué forma se podría calcular el sentimiento del usuario hacia un tema?
- ¿Es posible combinar los análisis de contenido y de sentimiento para cuantificar el Interés de un usuario en un tema?

1.6 Hipótesis

Describimos la hipótesis general a comprobar en esta tesis y una serie de hipótesis específicas que componen ésta. La hipótesis general que se comprueba en esta tesis es:

- Es posible definir formalmente el interés de un usuario en un tema y definir un procedimiento para cuantificarlo numéricamente

La cual esta compuesta por otras hipótesis que también se demuestran en el trabajo realizado en esta tesis:

- Es posible emplear técnicas de análisis de sentimiento en texto en español.
- Es posible aplicar minería de texto en español.

- Es posible combinar el análisis de sentimiento y la minería de texto en español.
- Los usuarios se pueden representar utilizando comentarios de sus contactos.
- El análisis de sentimiento se puede utilizar para crear representaciones del usuario.
- Los temas se pueden representar utilizando comentarios publicados en un microblog.
- Es posible cuantificar el interés de un usuario en un tema a partir de los resultados de la aplicación de análisis de sentimiento y minería de texto en español utilizando representaciones del usuario y del tema.

1.7 Contribución

Como mencionamos en la sección 1.1, el porcentaje de la población de usuarios que utilizan Internet en América Latina es del 42.9 %. El idioma Español ocupa el tercer lugar de los idiomas más utilizados en Internet según la empresa Miniwatts Marketing Group, la cual publica estadísticas acerca del uso de Internet [94]. Sin embargo, llama la atención que existen pocos recursos para trabajar con el idioma Español en el área de análisis de sentimiento— quizá debido a lo reciente del área. Por esta razón hemos decidio crear un recurso para hacer un análisis de sentimiento en español y que pueda ser utilizado por otros investigadores. Creamos un léxico de palabras con su polaridad de sentimiento (negativo o positivo) y una carga emotiva junto con otros componentes como lo es un conjunto de frases con su carga emotiva, y desarrollamos un procedimiento para evaluar el sentimiento de comentarios en español que utiliza este recurso para la evaluación. El léxico creado, sus componentes y el procedimiento de evaluación podrían ser utilizados por otros investigadores.

Para cuantificar el Interés tenemos que contestar las preguntas mencionadas en la Sección 1.5. Como parte de la contribución de esta tesis proponemos distintas

formas de representar a los usuarios para poder compararlos con los temas. Proponemos varias alternativas las cuales utilizan los comentarios de los contactos del usuario para modelar sus intereses, pero además filtramos dichos comentarios según su sentimiento, es decir, utilizamos sólo comentarios positivos, negativos, neutros o todos, para crear la representación y finalmente evaluamos la probabilidad de que dos usuarios sean amigos si tienen un Interés muy positivo o muy negativo sobre un tema.

Proponemos una forma de representación de los temas basada en la extracción de información proveniente del mismo ambiente donde los usuarios interactúan. Esta representación tiene la propiedad de ser dependiente del tiempo en que se crea, permitiendo obtener nuevas características del tema día con día; por ejemplo, los "Deportes" son un tema muy cambiante.

Desarrollamos un método para cuantificar el Interés del usuario en un tema utilizando análisis de sentimiento y minería de texto en el contenido generado por el usuario. Verificamos la validéz del modelo utilizando a los mismos usuarios, evaluando si la probabilidad de que sean amigos tiene alguna correlación con el Interés de ambos usuarios por un tema.

1.8 Estructura de la tesis

El resto de este trabajo está estructurado de la forma siguiente: En el capítulo 2 describimos el marco teórico en el que se basa nuestro modelo; hablamos de minería de texto, análisis de sentimiento y sistemas de recomendación. En el capítulo 3 describimos el modelo propuesto para la cuantificación del interés y la metodología que seguimos para crear una herramienta de análisis de sentimiento en español que fue necesaria para la experimentación. En el capítulo 4 describimos los experimentos realizados en cuanto al análisis de sentimiento y a la cuantifiación del interés en un tema. En este mismo capítulo introdujimos un posible procedimiento para hacer una recomendación de temas para un usuario con base a sus intereses utilizando

nuestro modelo. En el capítulo 5 comentamos las conclusiones de nuestro trabajo y mencionamos posibles trabajos que se podrían realizar para validar el modelo en otros ambientes, mejorar la herramienta TOM y posibles aplicaciones del modelo.

Capítulo 2

Marco Teórico

En este capítulo se describen los conceptos y teorías en las que se basa esta tesis. Se describen trabajos relacionados que tienen que ver con minería de texto y análisis de sentimiento, y que han inspirado la elaboración de las hipótesis aquí planteadas así como la metodología para obtener los resultados.

2.1 Análisis de sentimiento

El análisis de sentimiento es un área de investigación que se enfoca en las opiniones. Se trata primordialmente de un problema de clasificación de texto en el que se intenta saber si éste refleja una opinion positiva, negativa o neutra [63, 79, 64]. Para nuestros propósitos utilizaremos la definición de sentimiento propuesta por Alec Go, Richa Bhayani y Lei Huang [3], la cual se traduce al español como: "una sensación positiva o negativa de la persona". La Tabla 2.1 muestra algunos ejemplos.

El problema de clasificar el sentimiento en textos se ha abordado de diferentes formas; una de ellas es a través de la utilización de técnicas de aprendizaje de máquina supervisado como las máquinas de vectores de soporte (SVM) [80], Naïve Bayes [101]; aprendizaje de máquina no supervisado [102] o métodos estadísticos como Máxima Entropía [3]. Otra forma es mediante la creación manual o semi-automática de recursos informáticos como diccionarios de palabras o léxicos [98] y el uso de técnicas linguísticas basadas en conocimiento existente acerca del lenguaje y su estructura [27].

Tabla 2.1: Ejemplos de clasificación del sentimiento.

Sentimiento	Comentario
Positiva	la película me pareció muy buena
Negativa	el nuevo iPhone es muy caro
Neutra	acabo de pagar el seguro del auto

En esta tesis utilizamos un enfoque basado en léxicos y tomamos la decisión de crear uno nuevo para el idioma Español debido a que existen pocos recursos de este tipo y podría ser útil en trabajos futuros. El uso de éstos es común en el análisis de sentimiento para el idioma Inglés. La SentiWordNet 3.0 [4] es un diccionario de palabras (en inglés) creado semi-automáticamente a partir un conjunto de palabras descrito por Turney et al. [103] que contiene palabras consideradas "paradigmáticamente positivas" o "paradigmáticamente negativas". Éstas se agrupan y propagan sus valores de carga emotiva a otras palabras siguiendo enlaces como "además vea" o "vea antónimos" que existen en el diccionario WordNet 3.0 [36] publicado por la universidad de Princeton. A estas palabras se les asignan tres valores numéricos que dependen de la carga emotiva recibida indicando qué tan positiva, negativa o neutra es. Partiendo de una palabra y siguiendo estos enlaces se pueden ir visitando las palabras "vecinas" hasta una cierta profundidad k. Las cargas emotivas (positiva, negativa y neutra) son determinadas utilizando un comité clasificador que se compone de 8 miembros que resultan de la combinación de dos parámetros que son la profundiad k y un clasificador supervisado. Éstos corresponden a los valores de $k = 0, 2, 4, 6$ y los clasificadores Rocchio [86] y SVMs [107]. La idea que se utilizó para crear el diccionario es que si una palabra tiene muchos enlaces hacia palabras positivas (o negativas) es probable que también ésta sea positiva (o negativa).

Se han realizado trabajos utilizando estos diccionarios en otros idiomas traduciendo las palabras mediante un software especializado [25]. Sin embargo, en los microblogs se presentan muchas expresiones del lenguaje hablado que no tienen una traduccion directa y no son reconocidas por algún software, además de la gran can-

tidad de faltas de ortografía, uso indebido de palabras, lenguaje vulgar y lenguaje coloquial que hacen difícil realizar una traducción. Julian Brooke, Milan Tofiloski y Maite Toboada [16] reportan una pérdida de precisión al utilizar su herramienta llamada SO-CAL para analizar textos en español, adjudicando esta disminución al uso de una traducción de su diccionario. Esto nos motiva a crear un diccionario específico para el idioma Español.

Algunos trabajos también descomponen el texto en enunciados para procesarlos individualmente en busca de enunciados comparativos o para separar aquellos que contienen opiniones de los que no [53], [117], [115]. Otros han examinado la posibilidad de evaluar la opinión de un usuario sobre un tema utilizando información de sus contactos en una red social admás de los comentarios que éste publica en la misma [100]; a fin de cuentas, una empresa desea saber qué opinión tienen los usuarios con relación a un producto o servicio.

Como ya hemos mencionado, existen pocas herramientas para análisis de sentimiento de textos en español. Una de éstas es Sentitext [76], la cual ha mostrado tener un buen desempeño para determinar la polaridad del sentimiento [75]. Sentitext evalúa texto asignando una cantidad de "estrellas" que indican qué tan positivo o negativo es el sentimiento, entrega cero "estrellas" cuando el comentario es muy negativo, cinco "estrellas" cuando es neutro y diez "estrellas" cuando el comentario es muy positivo. Está basada en el uso de un diccionario de palabras y reglas del lenguaje. Utiliza un algoritmo que no implementa ninguna técnica de aprendizaje automático [76].

2.2 Minería de texto

La minería de texto es un proceso que consiste en extraer información útil de conjuntos de documentos no estructurados de texto, y en identificar automáticamente patrones interesantes no triviales o conocimiento [35, 34]. También puede ser visto como un problema de clasificación de texto [92] o de agrupación según ciertas simili-

tudes [9]. Se trata de un área multidisciplinaria que está fuertemente relacionada con otras: como la recuperación de información, análisis de texto, procesamiento de lenguaje natural, clustering, categorización (o clasificación), visualización, aprendizaje de máquina y minería de datos [46].

La recuperación de información se enfoca en la búsqueda de documentos en una colección grande de éstos que tentativamente contienen la respuesta a alguna necesidad de información [67]. Se diferencía del área de minería de texto en que ésta busca extraer información nueva, mientras que la recuperación de información busca optimizar el acceso a información a partir del uso de palabras clave en una consulta [2].

La minería de texto también puede ser vista como una extensión de la minería de datos en el sentido de que algunas técnicas utilizadas en ésta pueden ser aplicables en textos (tales como el clasificador Naïve Bayes, árboles de decisión, K vecinos más cercanos, máquinas de vectores de soporte y clustering) mediante un procesamiento previo del texto que lo permita [43]. La minería de datos es una disciplina que busca obtener información o conocimiento a partir de repositorios grandes de datos (incluyendo texto) [40]. Elizabeth D. Liddy [60] propone que la minería de datos hace referencia a un paso dentro de un área más general llamada Descubrimiento de Conocimiento en Datos (KDD, del inglés Knowledge Discovery from Data), el cual incluye en su definición los conceptos de almacenaje, acceso e interpretación mientras que la minería de datos hace referencia a la aplicación de algoritmos específicos para detectar y extraer patrones.

Feldman y Sanger dividen la minería de texto en cuatro áreas: tareas de pre-procesamiento, operaciones de minería, presentación y navegación, y técnicas de refinamiento [35].

Tareas de pre-procesamiento incluye todas aquellas rutinas, procesos y métodos requeridos para obtener y preparar los datos para las operaciones de descubrimiento de conocimiento.

Operaciones de minería incluye la detección de patrones, análisis de tendencias y algoritmos de descubrimiento de conocimiento.

Presentación y navegación incluye herramientas de visualización y navegación de los patrones orientadas al usuario.

Técnicas de refinamiento incluye los métodos que filtran información redundante, la ordenan, resumen, generalizan y agrupan.

En muchos casos el procesamiento de texto directamente desde el documento no es posible debido a las características del mismo, el cual es de naturaleza difusa, no estructurada y con ruido. Para lograr trabajar con documentos de manera más eficiente, usualmente se utilizan representaciones de los mismos que permitan trabajar con ellos con mayor facilidad. Comúnmente se hace mediante vectores de características (*feature vectors*) los cuales contienen características del documento necesarias para el procesamiento y su *peso* o valor de relevancia.

El modelo más común es el llamado "saco de palabras" (*bag-of-words*) el cual simplemente toma las palabras diferentes en el documento [43]. A cada palabra, posteriormente se le puede asignar un peso; lo cual se puede llevar a cabo de diferentes formas. La más simple es mediante el uso de un operador **binario** el cual equivale a 1 si la palabra está presente en el documento ó 0 si no lo está. La más utilizada es mediante el cálculo de valores TF-IDF[1][89] los cuales son valores que dan mayor peso a palabras poco frecuentes en todos los documentos y más frecuentes dentro de ciertos documentos en particular para los cuales serían más relevantes. El cálculo de este valor se realiza en esta tesis según la Ecuación 2.1.

$$\frac{f_{w,d}}{\sum_w f_{w,d}} \cdot \log\left(\frac{|D|}{f_{w,D}}\right) \tag{2.1}$$

Donde w representa la palabra, $f_{w,d}$ representa el número de veces que aparece la palabra w en el documento d, D es el conjunto de todos los documentos y $f_{w,D}$ es

[1] www.tfidf.com

el número de documentos en donde aparece la palabra w. Sin embargo, éste puede ser calculado de diversas formas. Manning, C.D. and Raghavan, P. y Schütze, H. [67] enlistan algunas variantes las cuales se muetran en la Tabla 2.2. En esta tesis utilizamos la Ecuación 2.1 para calcular el valor de relevancia de las palabras. Sin embargo, otras medidas también podrían ser utilizadas como en el método *binario*, que se basa en la simple existencia de la palabra en el documento. Otras medidas más complejas, que podrían estar en la categoría de aprendizaje supervisado son discutidas por Lan et al. [58] y Debole et al. [24].

Variantes de TF-IDF			
Frecuencia de Termino (TF)		Frecuencia de Documento (IDF)	
n (natural)	$tf_{w,d}$	n (no)	1
l (logaritmo)	$1 + log(tf_{w,d})$	t (idf)	$log\frac{N}{df_w}$
a (aumentado)	$0.5 + \frac{0.5tf_{w,d}}{max_w(tf_{w,d})}$	p (prob idf)	$max\{0, log\frac{N-df_w}{df_w}\}$
b (booleano)	$\begin{cases} 1 & \text{si } tf_{w,d} > 0 \\ 0 & \text{de otra forma} \end{cases}$		
L (log ave)	$\frac{1+log(tf_{w,d})}{1+log(ave_{w \in d}(tf_{w,d}))}$		

Tabla 2.2: Variantes del valor TF-IDF.

2.3 Sistemas de Recomendación

El área de sistemas de recomendación [1] ha adquirido un gran interés por la gran cantidad de aplicaciones prácticas y de información que existen en la actualidad. Estos sistemas tratan de aprender los gustos o preferencias de los usuarios con el fin de procesar la abundante información existente y ofrecerle recomendaciones [14]. Algunas aplicaciones de esto son: recomendación de música [114], películas [19], libros [44], artículos de tiendas virtuales como Amazon [62] y en general en el comercio electrónico [91, 90]. El problema de recomendación puede reducirse al problema de estimar calificaciones o *ratings* de un usuario hacia un artículo [1]; comúnmente, esto se hace con base en un historial de calificaciones otorgadas a artículos anteriormen-

te por el usuario. A este estimador también se le conoce como *función de utilidad*, que mide la utilidad de un artículo para un determinado usuario. Los sistemas de recomendación intentan encontrar los artículos que maximicen este valor para poder hacer la recomendación al usuario [1]. Balabanović y Shoham [6] clasifican a los sistemas de recomendación en tres categorías: recomendaciones basadas en contenidos (*content-based*), recomendaciones colaborativas (*collaborative*) y enfoques híbridos.

Recomendaciones basadas en contenido Al usuario se le hacen recomendaciones basadas en las preferencias que ha mostrado en el pasado.

Recomendaciones colaborativas Al usuario se le hacen recomendaciones de artículos que han sido del agrado de usuarios que tienen gustos similares a los de él.

Enfoques híbridos Estos métodos combinan las recomendaciones basadas en contenido y las colaborativas.

Los sistemas de recomendación tienen sus raíces en las investigaciones realizadas en el área de sistemas de recuperación de información [88, 5]. La mejora que incorporan los sistemas de recomendación es la presencia de *perfiles* de usuarios los cuales contienen información acerca de los gustos, preferencias o necesidades de éste [1]. Debido a esto la mayor cantidad de aplicaciones de los sistemas de recomendación existentes trabajan con texto y utilizan palabras clave o *keywords* para representar a los usuarios o artículos. Por ejemplo Pazzani y Billsus utilizan las 128 palabras más significativas para representar los documentos [81]. La medida de importancia de las palabras generalmente se calcula utilizando alguna variante del valor TF-IDF explicado en la sección anterior.

En los sistemas de recomendación muchas veces es necesario comparar usuarios para saber si tienen gustos similiares. Utilizando conjuntos de palabras clave más relevantes, una forma común de hacer esta comparación es mediante el uso de alguna heurística, como la similitud cosenoidal [88, 5]. La cual puede observarse en

la Ecuación 2.2; donde $\vec{w_1}$ y $\vec{w_2}$ son los vectores de los pesos de las palabras que representan a cada usuario.

$$Similitud(u_1, u_2) = \frac{\vec{w_1} \cdot \vec{w_2}}{\| \vec{w_1} \| \cdot \| \vec{w_2} \|} \tag{2.2}$$

Esta área de investigación permanece abierta y con una gran atención debido a las necesidades que surgen con el aumento de información disponible, sobre todo en Internet.

2.4 Trabajo relacionado

El problema de obtener información que satisfaga las necesidades de un usuario ha sido abordado desde los orígenes de la recuperación de información. En la medida que la tecnología ha avanzado y proliferado la cantidad de información disponible, las técnicas empleadas se han vuelto más complejas y han surgido otras áreas como los sistemas de recomendación y, más recientemente, el análisis de sentimiento. Actualmente existen pocos trabajos que integren análisis de sentimiento como parte de un método para obtener un modelo del usuario que represente sus intereses [59, 99, 47].

Tan et al. [100] crearon un modelo para predecir el sentimiento de un usuario con respecto a un tema basándose en los comentarios publicados por los usuarios y en las relaciones (contactos) cuyo sentimiento es conocido. Su enfoque es semi-automático ya que clasifican el sentimiento de un pequeño conjunto de usuarios manualmente y utilizan esa información para clasificar a los demás de forma automática. Matthew Michelson y Sofus a. Macskassy [71] crearon un métoto para obtener los temas de interés del usuario a partir de la selección de sustantivos en sus comentarios utilizando Wikipedia[2] como base de conocimientos para desambiguar los términos.

Jiang et al. diseñaron un modelo para clasificar el sentimiento de un comentario

[2] www.wikipedia.com

con relación a una entidad (*target*) [52]. Con este análisis intentan distinguir entre el sentimiento global del comentario y el sentimiento con respecto a dicha entidad o tema. Por ejemplo en el comentario "¿Porqué estoy recibiendo correo spam de gente extraña preguntándome si quiero chatear con **lady gaga**?". El sentimiento global podría ser negativo mientras que para Lady Gaga es un comentario neutro. Para representar a la entidad utilizan un método probabilístico con el cual extraen un conjunto de K palabras que se relacionan con ésta en mayor medida. Ellos utilizan el índice de Información Mutua Puntual (Pointwise Mutual Informacion, PMI) [31, 13]) para medir la relación entre dos palabras la cual se observa en la Ecuación 2.3.

$$PMI(w,t) = log\frac{p(w,t)}{p(w)p(t)} \tag{2.3}$$

Donde $p(w,t)$ es la probabilidad de que w y t ocurran en el mismo comentario y $p(w)$, $p(t)$ son las probabilidades de que w y t aparezcan en algún comentario. PMI es una medida de la asociación utilizada en estadística [68]. Las probabilidades son estimadas en un repositorio que contiene 20 millones de comentarios. En este trabajo ellos utilizan el valor de $K = 20$ para obtener las palabras que representan la entidad. Finalmente crean un grafo con los comentarios que contienen al menos una de estas 20 palabras el cual utilizan para evaluar su sentimiento.

Yi et al. evalúan el sentimiento de un documento (*review*) con respecto a un tema [113]. Utilizan técncias de Procesamiento de Lenguaje Natural (PLN) para obtener las palabras (*features*) más representativas en un conjunto de documentos relacionados con el tema, utilizando métodos matemáticos para calcular la relevancia de las mismas. Las palabras finalmente son seleccionadas por dos evaluadores humanos quienes las clasifican como *feature* o *no feature* y sólo las palabras en las que ambos evaluadores están de acuerdo en que son *feature*, se utilizan para la evaluación del documento. Finalmente, determinan el sentimiento del documento basándose en el sentimiento de cada enunciado encontrado que contenga alguna de las palabras elegidas. Nuevamente utilizando técnicas de PLN para determinar dónde comienza y termina un enunciado.

Malouf y Mullen [66] intentan descifrar los intereses políticos de los usuarios utilizando análisis de sentimiento, Naïve Bayes y clustering. Encontraron que utilizando solamente el texto publicado por el usuario es difícil determinar estos intereses, y que la precisión se incrementa cuando se utiliza la ubicación del usuario en la red social como información. Utilizaron el modelo PMI-IR de Turney [102] para analizar el sentimiento, el cual logró una precisión de 40.76 %. Utilizando Naïve Bayes alcanzaron un 63.59 % y utilizando clustering obtuvieron un 68.48 % de precisión. Soo-Min y Hovy introducen el problema de encontrar sentimientos en un texto sobre un tema y a las personas que los expresan [55]. Utiliza PLN para encontrar enunciados que contienen referencias al tema y a quién expresa el sentimeinto. El analizador de sentimiento que utilizan fue desarrollado por ellos mismos utilizando una lista de palabras como semilla y aumentándolas utilizando WordNet [72]. El sentimiento de cada oración se determina mediante dicho analizador. Mei et al. determinan el sentimiento de un documento con respecto a un tema basandose en un método probabilístico llamado Latent Semantic Indexing (LSI) [42] para asignar *scores* a cada enunciado en el documento según su probabilidad de pertenecer a un tema y de tener un sentimiento [70]. Finalmente el sentimento existente en el documento se calcula mediante una fórmula. Otros trabajos similares utilizan Latent Dirichlet Allocation (LDA) para obtener los temas y los sentimientos en un documento automáticamente [61, 119].

Wang et al. incorporan análisis de sentimiento en un sistema de recomendación basado en filtrado colaborativo, en el cual los usuarios pueden asignar etiquetas (*tags*) tales como ("divertida", "aburrida", etc.) a películas [110]. El análisis de sentimiento se lleva a cabo en dichas etiquetas generando así nueva información que incrementa la utilidad de las recomendaciones alcanzando una precisión de 29.70 % que sobrepasa la precisión de 25.12 % y 27.31 % en comparación con métodos que no utilizan información acerca del sentimiento de las etiquetas. Se basan en el trabajo de Durao y Dolog el cual utiliza métricas clásicas de la recuperación de información como TF-IDF y similitud cosenoidal, pero agrega un componente de análisis de sentimiento [29].

2.5 Caso de estudio: Twitter

En esta tesis trabajamos con el microblog Twitter ya que es uno de los más populares actualmente en México y Estados Unidos. Twitter (http://www.twitter.com), es un microblog que se basa en la propagación de pequeños comentarios llamados *tweets*, los cuales son publicados por los usuarios y distribuidos automáticamente a otros usuarios que deseen recibir actualizaciones de éstos. Cada comentario está limitado a 140 caracteres de longitud y diariamente se publican al rededor de 200 millones [105], los cuales pueden ser vistos por cualquier usuario registrado. Estas características lo hacen interesante para el análisis de sentimiento debido a que los usuarios tienden a ir directamente al punto de lo que quieren decir, los mensajes pueden ser procesados rápidamente por los algoritmos y es relativamente fácil obtener una cantidad suficiente de comentarios para un análisis.

2.5.1 Investigaciones que han utilizado a Twitter

Algunos autores han comenzado recientemente a utilizar este microblog para sus investigaciones. Twitter cuenta con un conjunto de utilerías que se ofrecen gratuitamente por la compañía y que permiten obtener una cantidad de información suficiente para el estudio científico. Entre algunos trabajos podemos mencionar a la herramienta Twitter-Sentiment [3, 37] para el análisis de sentimiento. Twitter-Sentiment se basa en la evaluación de comentarios por separado (independientemente del contexto) y los clasifica como negativos, neutros o positivos según su carga emotiva. Esta herramienta no utiliza información de los contactos del usuario ni intenta determinar su posición con respecto a un tema sino que se limita a evaluar el comentario en sí mismo.

Jonathon Read explica cómo se pueden utilizar emoticonos para crear conjuntos de prueba extraídos de Twitter que sirvan para el entrenamiento de modelos de aprendizaje automático en la clasificación del sentimiento [85]. En la televisión cada

vez es más frecuente el uso de redes sociales para tener contacto con los televidentes. Nicolas A. Diakopoulos y David A. Shamma [26] realizaron un estudio acerca de los cambios en el sentimiento en los comentarios publicados en Twitter con relación al debate que se llevaba a cabo entre los candidatos presidenciales Barack Obama y John McCain en vivo, en Estados Unidos. Otro estudio similar es el realizado por Brendan O'connor y cols. en el que encontraron que el sentimiento en los comentarios publicados en Twitter mantiene una correlación positiva con los resultados de encuestas publicados por las revistas Index of Consumer Sentiment (ICS) para la confianza de la gente en la economía del país y Gallup[3] para la preferencia de la gente por los mismos candidatos presidenciales en el 2008 [74]. Bernard et al. hicieron un estudio del efecto que tiene la publicidad de boca en boca dentro del microblog y concluyen diciendo que la tendencia de esta red a ser utilizada como una fuente confiable de información representa una nueva oportunidad para contactar clientes potenciales [49]. Cha et al. realizaron un estudio sobre la influencia de ciertos usuarios en el comportamiento de los demás; encontraron que no necesariamente aquellos con la mayor cantidad de seguidores son los que tienen mayor poder para influir, y que este poder tampoco es obtenido sin esfuerzo sino mediante una participacion activa [18]. Otro tipo de estudios exploran las características de comunidades o usuarios particulares en Twitter a través de sus comentarios e intentan descubrir información como: ¿De qué temas hablan?, ¿Están a favor o en contra de algo? o ¿Como se relacionan entre sí? [100, 51, 45, 118].

[3]www.gallup.com/poll/111439/obama-mccain-two-bestliked-candidates.aspx

Capítulo 3

Cuantificación del Interés

En este capítulo explicamos el modelo conceptual para cuantificar el interés de un usuario en un tema. El procedimiento se realiza en tres etapas por lo que comenzamos desde un punto de vista general mediante un esquema y continuaremos hasta considerar todos los detalles del método.

3.1 Definición del Interés

El interés es una medida del agrado, desagrado o indiferencia del usuario hacia un tema. Cuando un usuario habla acerca de algo, puede utilizar palabras que llevan una carga emotiva ya sea positiva, negativa o neutra, y éstas nos permiten conocer la opinión de ese usuario. Si el usuario menciona el tema en forma repetitiva, esto también es indicativo de un cierto interés por parte del usuario aunque lo haga de manera objetiva sin expresar sentimientos.

Cuantificamos el interés contestando primeramente la pregunta ¿Cuánto habla el usuario acerca de un tema?, la cual nos permite saber que tan común es que el usuario hable del tema, si lo hace cotidianamente o si por el contrario lo hace poco. Después contestamos la pregunta ¿Qué sentimiento utiliza el usuario al hablar del tema? que nos permite saber la opinión del usuario con relación al tema. Si un usuario habla bien o positivamente acerca de un tema decimos que tiene interés en él, pero cuando habla mal o negativamente, aunque también tiene un interés, éste es un interés negativo que indica desagrado por parte del usuario.

El interés del usuario es calculado mediante la combinación de la respuesta a estas dos preguntas a través de una función a la que llamamos *función de cuantificación*. En la Definición 1 describimos el Interés con mayor formalidad.

Definición 1. *El Interés I de un usuario u con respecto a un tema t es una función del contenido generado por el usuario con respecto al tema $c_u(t)$ y del sentimiento expresado con respecto al tema $s_u(t)$. Es un número real cuyo signo y magnitud indican el nivel de agrado ($I > 0$), desagrado ($I < 0$) e indiferencia ($I \approx 0$) con respecto a un tema. El Interés podría ser positivo, negativo o neutro dependiendo del tema y se muestra en la Ecuación 3.1. Donde f es la función de cuantificación.*

$$I_u(t) = f(c_u(t), s_u(t)) \tag{3.1}$$

Definimos, entonces, el Interés en un tema como una función de similitud de contenido generado por el usuario y del análisis de sentimiento aplicado en dicho contenido. La similitud de contenido $c_u(t)$ evalúa el parecido que existe entre el contenido del usuario y el tema. Es un número real que se encuentra en el rango [0, 1], el cual toma el valor de 0 cuando éstos son totalmente diferentes y 1 cuando son exactamente iguales. Esta función contesta la pregunta ¿Cuánto habla el usuario acerca de un tema?. El análisis de sentimiento $s_u(t)$ evalúa el sentimiento del usuario hacia el tema. Es un número real que se encuentra en el rango $[-1, 1]$, el cual toma el valor de -1 cuando todos los comentarios del usuario referentes al tema son negativos, el valor de 1 cuando son todos positivos y el valor de 0 cuando son todos neutros. Esta función contesta la pregunta ¿Qué sentimiento utiliza el usuario al hablar del tema?.

La *función de cuantificación* utiliza las métricas de similitud de contenido y sentimiento para cuantificar el interés. Ésta podría ser cualquier función elegida de forma que pueda cumplir con su objetivo de hacer estimaciones con relación al interés del usuario. Por simplicidad utilizamos la operación de multiplicación para definir esta función. La multiplicación respeta el signo del sentimiento y, para magnitudes

pequeñas (o grandes) de las funciones $s_u(t)$ y $c_u(t)$, el interés será pequeño (o grande) también. Utilizando la multiplicación como *función de cuantificación* podemos escribir el Interés como en la Ecuación 3.2.

$$I_u(t) = c_u(t) \cdot s_u(t) \tag{3.2}$$

Dado que utilizamos la multiplicación como función de cuantificación, el Interés es un número real que se encuentra en el rango de [-1, 1] donde valores negativos indican un desagrado por parte del usuario, valores positivos indican un agrado y valores cercanos a cero indican indiferencia.

En lo subsiguiente, al hablar del Interés hacemos referencia a la Ecuación 3.2, la cual utilizamos para realizar los experimentos que comprueban nuestras hipótesis.

3.2 Descripción del proceso de cuantificación del Interés

La cuantificación del interés del usuario en un tema se realiza en tres pasos según el modelo propuesto. El primero de ellos calcula la similitud que existe entre el contenido generado por el usuario y el tema, el segundo paso evalúa el sentimiento que el usuario expresa y el tercero cuantifica el Interés mutiplicando ambas métricas. En la Figura 3.1 se muestra una representación conceptual del modelo.

El contenido generado por el usuario y el tema funcionan como entradas para el proceso de cuantificación que finalmente arrojará el Interés. Ambas funciones $c_u(t)$ y $s_u(t)$ utilizan información proveniente de estas dos fuentes para hacer los cálculos y enviar los resultados a la función de cuantificación. Para poder evaluar $c_u(t)$ y $s_u(t)$ es necesario utilizar representaciones del usuario y del tema que permitan hacer comparaciones entre ellos. Nosotros utilizamos vectores de valores TF-IDF de palabras para representar tanto al usuario como al tema ya que éstos permiten

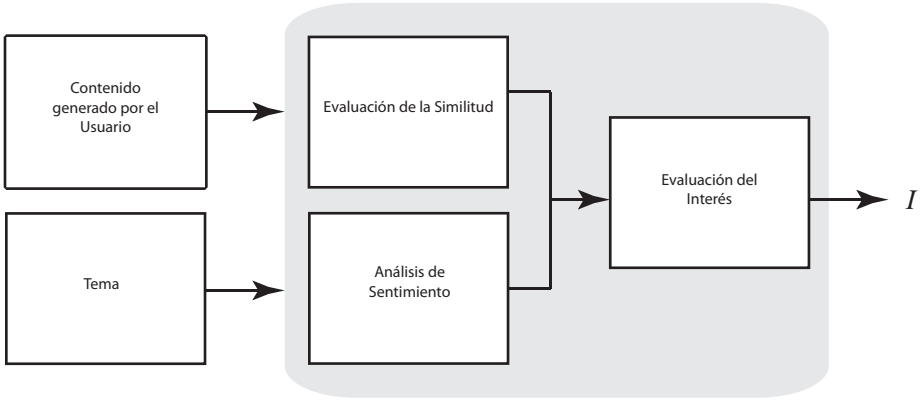


Figura 3.1: Proceso de cuantificación del Interés de un usuario en un tema.

realizar operaciones matemáticas como la similitud cosenoidal. Esto será explicado en las siguientes secciones. El análisis de sentimiento se lleva a cabo utilizando una herramienta de elaboración propia llamada TOM (Twitter Opinion Mining) la cual será explicada a detalle más adelante. Con ayuda de TOM se puede clasificar el sentimiento global del usuario hacia el tema. El Interés es calculado finalmente utilizando la Ecuación 3.2.

3.3 Representación del usuario y del tema

En los microblogs, los usuarios generan contenido a partir de pequeños comentarios por medio de los cuales buscan compartir algún pensamiento o información. Nosotros utilizamos estos comentarios para crear representaciones tanto de los usuarios como de los temas. La diferencia entre una representación y otra está en la forma en que se seleccionan los comentarios. A partir de los comentarios seleccionados creamos un vector que representará al usuario o al tema, cuya dimensión es el número de palabras diferentes en los comentarios y sus componentes son los valores TF-IDF de las palabras. El valor TF-IDF (del inglés, *Term Frecuency - Inverse Document Frecuency*) representa la relevancia de una palabra dentro de un documento y es utilizado con frecuencia en el área de recuperación de información [41]. Tiene la característica de incrementar su valor para palabras que se repiten dentro del

$$\begin{pmatrix} Lanza \\ Android \\ su \\ Appstore \\ ni \\ mas \\ menos \\ que \\ en \\ Amazon \end{pmatrix} \rightarrow \begin{pmatrix} \frac{1}{11} \cdot log(\frac{3}{2}) = 0.01600 \\ \frac{1}{11} \cdot log(\frac{3}{1}) = 0.04337 \\ \frac{1}{11} \cdot log(\frac{3}{1}) = 0.04337 \\ \frac{1}{11} \cdot log(\frac{3}{1}) = 0.04337 \\ \frac{2}{11} \cdot log(\frac{3}{1}) = 0.08674 \\ \frac{1}{11} \cdot log(\frac{3}{1}) = 0.04337 \\ \frac{1}{11} \cdot log(\frac{3}{1}) = 0.04337 \\ \frac{1}{11} \cdot log(\frac{3}{2}) = 0.01600 \\ \frac{1}{11} \cdot log(\frac{3}{1}) = 0.04337 \\ \frac{1}{11} \cdot log(\frac{3}{1}) = 0.04337 \end{pmatrix}$$

Figura 3.2: Ejemplo de obtención de valores TF-IDF para un conjunto de palabras de un documento.

documento y que no son muy comunes en el repositorio de documentos en general. Cuando una palabra aparece en varios documentos del repositorio o si se utiliza con poca frecuencia dentro de un documento pierde relevancia para dicho documento.

Para ilustrar mejor el efecto que tiene el utilizar valores TF-IDF, consideremos el siguiente ejemplo que contiene tres documentos los cuales estan formados por un solo comentario.

1. Es probable que termine con un iPhone 4S antes de que termine el año.
2. Telmex lanza tienda de aplicaciones.
3. Lanza Android su Appstore ni mas ni menos que en Amazon.

Las palabras que componen el documento 3 son: “Lanza”, “Android”, “su”, “Appstore”, “ni”, “mas”, “ni”, “menos”, “que”, “en”, “Amazon”. Utilizando la Ecuación 2.1 para calcular el valor TF-IDF de estas palabras en el repositorio de tres documentos, obtenemos los valores de la Figura 3.2.

Nótese que la palabra "ni" aparece una sola vez en el vector ya que por definición el vector contiene solo palabras diferentes. Al utilizar TF-IDF, la palabra "Lanza" obtiene menor relevancia ya que está presente en dos documentos, en cambio la palabra "ni" adquiere mayor relevancia porque se repite dos veces en el documento y no aparece en los demás. Aunque el valor TF-IDF será muy pequeño para ciertas palabras que se utilizan mucho como "la" o "el" es conveniente eliminar previamente palabras que de antemano sabemos que serán poco relevantes y que por lo tanto serán poco útiles para cuantificar el interés. A estas palabras se les conoce como "vacías" o "stopwords" en inglés. Nosotros utilizamos una lista de 308 palabras que forman parte del mecanismo de indexación de Lucene. Estas palabras son excluídas. Entre algunas de éstas podemos mencionar: "la", "de", "ya", "esta", "sin", "había", "tiene", "sea", "eran", "poco", "algo", "ni", "él", "muchos", etc.

En las subsecciones siguientes describiremos con mayor detalle estas representaciones.

3.3.1 Representación del usuario

La representación de los usuarios se crea a partir de los comentarios de sus contactos. Los contactos fueron útiles para crear un sistema de recomendación de noticias llamado Buzzer [82] el cual utiliza los comentarios de éstos para ordernar por relevancia las noticias y así poder hacer la recomendación al usuario. En esta estrategia ellos excluyeron los comentarios del propio usuario. Es común en algunos microblogs, que algunos usuarios tienden a no publicar comentarios o hacerlo muy poco; sin embargo, tienen contactos que utilizan como fuentes de información [51]. Por esta razón damos preferencia al uso de los contactos para representar al usuario.

En la Definición 2 exponemos el concepto de **vector de usuario** que estaremos utilizando en esta tesis. Como mencionamos en la Sección 3.2, los comentarios son clasificados según su sentimiento en positivos, negativos y neutros por la herramienta TOM. Esto permite que podamos utilizar la clasificación para crear subconjuntos

de comentarios que tienen el mismo sentimiento y utilizar éstos para representar al usuario. Como parte de la experimentación analizamos el efecto que tiene en la cuantificación del interés el utilizar solo comentarios positivos, negativos o neutros provenientes de los contactos del usuario.

Definición 2. *El* ***vector de usuario*** $\vec{U}$ *es un vector de dimensión* N*, donde* N *es el número total de palabras diferentes que existen en los comentarios de sus contactos y sus componentes son los valores TF-IDF de cada palabra.*

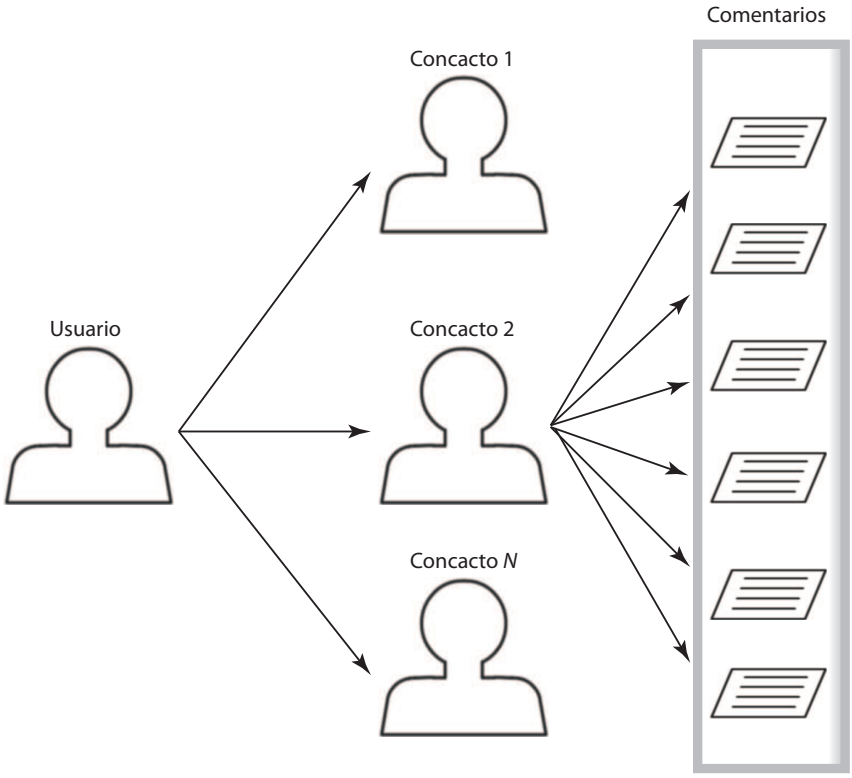


Figura 3.3: Representación del usuario en base a los comentarios de sus contactos.

En la Figura 3.3 se muestra la forma en que son seleccionados los comentarios. De cada uno de los contactos del usuario tomamos los k comentarios más recientes. Consideramos que un valor adecuado para k es 1,000; dicho valor fue utilizado para los experimentos que se describirán en el Capítulo 4.

3.3.2 Representación del tema

Definición 3. *El **vector de tema** $\vec{T}$ es un vector de dimensión N, donde N es el número total de palabras diferentes que existen en los comentarios elegidos para representar el tema y sus componentes son los valores TF-IDF de cada palabra.*

Para obtener los comentarios que se utilizarán para crear el *vector de tema* hacemos una búsqueda de todos aquellos comentarios que incluyan un conjunto determinado de palabras sin importar el usuario que lo haya publicado. Estas palabras se eligen dependiendo del tema. Por ejemplo, al utilizar el tema de Deportes, utilizamos las palabras "deportes" y "deporte". En esta primera aproximación utilizamos el nombre del tema junto con algunas variantes como el plural o singlar de las palabras para buscar aquellos comentarios que contengan alguna de ellas. Jiang et al. utilizan un enfoque parecido para buscar comentarios acerca de una entidad (o *target*) [52]. Ellos utilizan las palabras "Obama", "Google", "iPad", "Lakers" y "Lady Gaga" para obtener conjuntos de comentarios que tengan un sentimiento hacia esas entidades.

3.4 ¿Cuánto habla el usuario acerca de un tema?

Para dar respuesta a la pregunta ¿Cuánto habla el usuario acerca de un tema? calculamos la similitud cosenoidal entre el *vector de usuario* $\vec{U}$ y el *vector de tema* $\vec{T}$. La similitud cosenoidal es una medida de la similitud entre dos vectores que mide el ángulo θ existente entre ellos [5] y se calcula de la siguiente forma:

$$Similitud(\vec{U},\vec{T}) = cos(\theta) = \frac{\vec{U}\cdot\vec{T}}{\parallel \vec{U} \parallel \parallel \vec{T} \parallel} = \frac{\sum_{j=1}^{t} x_j \cdot y_j}{\sqrt{\sum_{j=1}^{t} x_j^2 \cdot \sum_{j=1}^{t} y_j^2}} \tag{3.3}$$

Ésta da como resultado un 0 cuando los vectores son totalmente diferentes, es decir, los usuarios no han utilizado ninguna palabra en común, lo cual es poco probable, y da como resultado un 1 cuando los vectores son idénticos. Si sustituímos la similitud cosenoidal en la fórmula del Interés quedaría de la siguiente forma:

$$I_u(t) = Similitud(\vec{U}, \vec{T}) \cdot s_u(t). \tag{3.4}$$

La similitud cosenoidal es utilizada con frecuencia en el área de recuperación de información y es utilizada por Zhong Wang et al [110] en un sistema de recomendación de películas.

3.5 ¿Qué sentimiento utiliza el usuario al hablar del tema?

El cálculo del sentimiento que el usuario utiliza al hablar de un cierto tema se realiza mediante la obtención de un promedio del sentimiento encontrado en los comentarios del usuario que contienen ciertas palabras clave que dependen del tema en cuestión. De cada tema, elegimos N palabras que tienen los valores TF-IDF más altos, es decir, las palabras más releventas del tema. Para cada una de las palabras se obtiene una lista de comentarios de los contactos del usuario que la contienen y se calcula un promedio del sentimiento.

La evaluación de los comentarios se realiza utilizando la herramienta TOM. Esta herramienta clasifica los comentarios según su sentimiento en positivos, negativos o neutros. Para obtener el promedio del sentimiento convertimos esta clasificación a un número entero según la Tabla 3.1. El sentimiento final del usuario con respecto al tema se calcula obteniendo un promedio del sentimiento con respecto a las palabras individuales según la Ecuación 3.5. De esta forma podemos obtener el valor de $s_u(t)$ en la función del Interés, dando una respuesta a la pregunta ¿Qué sentimiento utiliza el usuario al hablar del tema?. A continuación definimos el sentimiento del usuario

con respecto a un tema.

Definición 4. *El sentimiento del usuario con respecto a un tema es el promedio del sentimiento individual con respecto a las palabras más relevantes del tema. Sea T_N el conjunto que contiene las N palabras más relevantes del tema (con los valores TF-IDF más grandes), se sigue que:*

$$s_u(t) = \frac{C_{pos} - C_{neg}}{C_{pos} + C_{neg} + C_{neu}} \tag{3.5}$$

Donde $C_{pos}, C_{neg}, C_{neu}$ son la cantidad de comentarios positivos, negativos o neutros respectivamente, que contienen alguna palabra $w \in T_N$.

En esta tesis trabajamos con el valor $N = 400$ ya que es una cantidad estadísticamente significativa.

Clasificación	Número
Positivo	1
Neutro	0
Negativo	-1

Tabla 3.1: Conversión de clasificación de sentimiento a número entero.

3.6 TOM: Twitter Opinion Mining

La evaluación del sentimiento de los comentarios de los usuarios se lleva a cabo utilizando una herramienta de análisis de sentimiento de elaboración propia. Esta herramienta está basada en un léxico de palabras clasificadas como positivas o negativas y con un peso según la fuerza del sentimiento. Así mismo contiene un diccionario de frases con clasificación y peso e implementa intensificadores de valencia en su análisis. Al conjunto de palabras, frases e intensificadores de valencia lo llamamos *diccionario*. El *diccionario* y su creación se describe a continuación y más adelante describiremos el algoritmo que utiliza TOM para la clasificación.

3.6.1 Diccionario

Llamamos *diccionario* a los recursos informáticos que utiliza TOM para clasificar el sentimiento de los comentarios. El diccionario tiene tres componentes que se definen a continuación.

Léxico: Es un conjunto grande de palabras con un sentimiento positivo o negativo y un peso según la fuerza del sentimiento. Los léxicos se han utilizado ampliamente en la investigación del análisis de sentimiento. El léxico utilizado por TOM fue creado de forma semi-automática mediante de la traducción de un léxico existente en inglés y su enriquecimiento a través de la obtención de nuevas palabras utilizando frases que permitieran extraerlas semti-automáticamente. El léxico es utilizado para clasificar palabras individuales dentro del comentario.

Frases: Es un conjunto de frases comunes del lenguaje con un peso sentimental. Las frases son una ampliación al léxico ya que incluyen un conjunto de palabras que deben ser escritas en el orden correcto para ser identificadas dentro del comentario. Una frase puede cambiar por completo el significado de las palabras individuales y esto hace que el reconocimiento de las frases sea un paso necesario para lograr un alto nivel de precisión en la clasificación. Por ejemplo, la frase “vale la pena” que se encuentra en el texto “la nueva pantalla táctil del iPhone vale la pena” modifica por completo el valor de la palabra individual “pena” que por sí sola es considerada negativa y en cambio, al usarse en esta frase cambia el valor del sentimiento a positivo.

Modificadores de valencia: Es un conjunto de palabras o expresiones que sirven para enfatizar lo que se está diciendo. Éstos pueden cumplir cuatro funciones: aumentar, disminuir, invertir o neutralizar el sentimiento. En la Tabla 3.2 se muestran ejemplos de cómo se utilizan estos modificadores de valencia. Los modificadores de valencia se pueden aplicar a otros modificadores de valencia como en la frase “muy muy interesante” o para modificar el valor de una frase como en “no vale la pena” donde el “no” hace la función de invertir el valor positivo y volverlo negativo. Para

Tabla 3.2: Uso de modificadores de valencia.

Modificador	Ejemplo	Efecto
realmente	La película fue realmente buena	Aumentar
muy	El nuevo iPhone es muy bonito	Aumentar
algo	El mensaje fue algo positivo	Disminuir
poquito	La ensalada quedó un poquito desabrida	Disminuir
no	no vale la pena seguir así	Invertir
no	no es tan malo como algunos pensaban	Neutralizar

tener un entendimiento más amplio de los modificadores de valencia recomendamos leer a Polanyi y Zaenen [83] quienes los describen con más detalle. Por ahora nuestro diccionario cuenta con un cojunto limitado de estos modificadores ya que implementa sólo algunos de los intensificadores mencionados en el trabajo citado.

3.6.2 Creación del diccionario

En esta sección describimos el procedimiento seguido para crear el diccionario que utiliza TOM. Para la creación de las partes del diccionario utilizamos una colección extraída de Twitter de 6,000 comentarios publicados por usuarios diferentes, esto es, cada comentario proviene de un usuario diferente. Los comentarios fueron divididos en dos grupos de 3,000 comentarios que llamaremos C_1 y C_2 para hacer referencia a ellos más adelante. Cada uno de los comentarios en ambos conjuntos fue clasificado manualmente por un grupo de tres personas, cada una en lugar y tiempo diferente de las demás. Dos de las personas que evaluaron los comentarios son estudiantes de Maestría y una Doctora. La evaluación se realizó a través de Internet en donde los evaluadores podían observar y clasificar los comentarios (ver Figura 4.1). Los evaluadores no tuvieron comunicación entre ellos durante el tiempo en que clasificaron los comentarios y no se definió ningún tipo de criterio para decidir cuando un comentario sería positivo, negativo o neutro. Esto permitió que

los evaluadores pudieran clasificar libremente, sin ningun tipo de prejuicio en cuanto al sentimiento. El conjunto C_1 fue utilizado para la creación del léxico, conjunto de frases y modificadores de valencia que describiremos en las secciones siguientes y el conjunto C_2 fue apartado para la experimentación y validación de TOM.

Creación del léxico

La creación del diccionario se dividió en cuatro etapas. La primera consistió en la traducción del conjunto de palabras publicado por Liu [10] mediante el traductor inglés a español de Google[1]; la segunda etapa consitió en agregar palabras y frases encontradas en los comentarios; la tercer etapa consitió en la expansión del diccionario para incluir palabras nuevas a partir de las que ya se tenían y la cuarta etapa consistió en asignar un valor que representa la fuerza (peso) del sentimiento. A continuación detallaremos cada una de las etapas.

El diccionario publicado por Liu[2] contiene 6,800 palabras clasificadas como positivas o negativas. Se utilizó el traductor de Google para traducir estas palabras a español; se limpiaron manualmente ambos conjuntos de palabras eliminando aquellas que se repetían debido a la traducción y reclasificamos aquellas que al traducirse automáticamente cambiaron su sentimiento. Estas palabras fueron introducidas al diccionario en la primer etapa.

En la segunda etapa nos basamos en el uso de emoticonos al igual que Read [85] para extraer comentarios de Twitter. Los emoticonos que utilizamos se muestran en la Tabla 3.3. Se realizó una búsqueda en los comentarios para obtener aquellos que contuvieran cualquiera de los emoticonos positivos o negativos y se separaron según su categoría. Obtuvimos 2.2 millones de comentarios que contienen emoticonos positivos y 0.4 millones de comentarios con emoticonos negativos. Esto es congruente con los resultados de Rodríguez y Torres [87] en los que se observa una tendencia a escribir más comentarios positivos que negativos. Para nuestros propósitos, supo-

[1] translate.google.com

[2] http://www.cs.uic.edu/ liub/FBS/sentiment-analysis.html

Tabla 3.3: Emoticonos utilizados para la separación de comentarios.

Positivos	Negativos
:)	:(
:-)	:-(
:D	:@
:-D	:-@

Tabla 3.4: Estructuras utilizadas para obtener palabras de sentimiento.

Primera	Segunda	Ejemplo
es	un	sabio
eres	una	bailarina
son	unas	tremendas
son	unos	listillos
estas	bien	triste
esta	bien	bonita

nemos que en estos comentarios el usuario está expresando una sensación positiva o negativa dada la presencia explícita de emoticonos. Utilizamos los comentarios para crear una lista de todos los trigramas (conjuntos de tres palabras) que aparecen en ellos y se seleccionaron aquellos que siguieron el patron "es-un" o alguna de sus variantes (ver la Tabla 3.4), almacenando la tercer palabra como una palabra que expresa un sentimiento. Se obtuvieron 6,145 palabras, las cuales fueron clasificadas manualmente como positivas, negativas o neutras. Las palabras clasificadas como neutras fueron desechadas y se incorporaron al diccionario sólo aquellas clasificadas como positivas o negativas. Se obtuvo un total de 1,065 palabras positivas y 1,147 palabras negativas.

En Monterrey, el uso de palabras en inglés durante conversaciones o en las publicaciones dentro de redes sociales es una práctica común. Ciertas palabras como "nice", "good", "cool", "perfect" entre otras, son necesarias para la correcta clasificación de comentarios aún cuando se traten de comentarios en español.

Tabla 3.5: Criterio de asignación de pesos.

Peso	Descripción	Ejemplo
+3	Sensación positiva + Halago + Alabanza	Es excelente en su trabajo
+2	Sensación positiva + Halago	Eres buen músico
+1	Sensación positiva	Obtuvo un premio
-1	Sensación negativa	Tuvo un accidente
-2	Sensación negativa + Insulto	Lo despidieron por corrupto
-3	Sensación negativa + Insulto + Humillación	Tiene una actitud mediocre

En la tercer etapa decidimos expandir el diccionario agregando palabras nuevas y asignándoles una carga emotiva. Para esto utilizamos la herramienta Freeling[3] [78] la cual permite hacer un análisis del lenguaje en español, es gratuita y se ha utilizado en muchos trabajos de investigación [16]. Utilizamos esta herramienta para detectar todos los verbos que se encontraban en nuestro diccionario y posteriormente los conjugamos.

En la cuarta etapa se recorrió el diccionario asignando valores de carga emotiva a cada palabra. Está asignación se realizó manualmente. Los valores que se eligieron variaron del -3 al +3 para el más negativo y el más positivo respectivamente. El criterio que se empleó para la asignación de las palabras se muestra en la Tabla 3.5. A las palabras que resultaron de la conjugación de los verbos se les asignó la misma carga emotiva que la palabra original.

Creación del conjunto de Frases

El conjunto de frases fue creado seleccionando frases que se encontraran en el conjunto de comentarios C_1 mediante inspección visual. Frases como "vale la pena", "beneficio de la duda", "envidia de la buena", fueron incorporadas en nuestro diccionario. En estas frases de ejemplo se observa que una valoración palabra por palabra resultaría en una clasificación incorrecta. Por ejemplo, la palabra "envidia",

[3]http://nlp.lsi.upc.edu/freeling/

por sí sola es negativa y al escribirse dentro de la frase "envidia de la buena" se invierte su polaridad. Algunas otras frases como "ni trabajan ni dejan trabajar" pueden generalizarse. Para este tipo de frases se utilizó una colección de verbos extraída de la base de datos de Freeling [78] de manera que el algoritmo detecta patrones como "ni [verbo] ni dejan [verbo]". A cada frase se le asignó manualmente un peso que representa su carga emotiva que, al igual que las palabras, se encuentra en un rango de -3 a +3.

Creación del conjunto de Modificadores de Valencia

Para crear el conjunto de modificadores de valencia seguimos el mismo proceso que para crear nuestro conjunto de frases. Mediante una inspección visual del conjunto de comentarios C_1 obtuvimos una lista de modificadores de valencia comunes que incorporamos al diccionario.

3.6.3 Diseño del algoritmo

El algoritmo de clasificación de sentimiento está compuesto por varias etapas que generan información con respecto al comentario. Los pasos en del algoritmo son los siguientes:

1. División del comentario en palabras
2. Corrección de ortografía
3. Procesamiento de léxico
4. Procesamiento de frases
5. Procesamiento de preguntas
6. Procesamiento de modificadores de valencia
7. Evaluación final

División del comentario en palabras

El primer paso del algoritmo es dividir el comentario en palabras. El proceso de dividir es sencillo: cualquier caracter que no sea una letra de la “a” a la “z” se considera un delimitador que marca el final de la palabra actual y comienza una nueva. Cuando de nuevo se encuentra con una letra de la “a” a la “z” vuelve a comenzar una nueva palabra. Así, la frase “@usuario JB con Pwntools 3.0 funcionando de maravilla en un iphone 2g” quedaría dividida de la siguiente forma: {“@”, “usuario”, “JB”, “con”, “Pwntools”, “3.0”, “funcionando”, “de”, “maravilla”, “en”, “un”, “iphone”, “2”, “g”}.

Corrección de ortografía

A cada palabra en el arreglo se le aplica un filtro que elimina caracteres repetidos que provcarían que la palabra no fuera encontrada en el léxico. Por ejemplo, la palabra “excelenteeeee” será transformada a la palabra “excelente”. Este tipo de expresiones son comunes, en los microblogs por lo que es necesario realizar este paso.

Procesamiento de léxico

Una vez terminadas estas preparaciones se procede a buscar palabras del comentario en el léxico. Cada palabra encontrada se marca con una carga emotiva que es el peso asignado a la palabra en el léxico; en caso de no encontrarse, se considera neutra y se asigna el valor 0.

Procesamiento de frases

Al igual que con las palabras del léxico, se buscan frases (conjuntos de palabras) en el comentario que existan en nuestro conjunto de frases y se les asigna una carga emotiva (el peso almacenado para dicha frase). Cuando se encuentra una frase en el comentario, ésta se agrupa y a partir de ese momento se trata como si fuera una

sola palabra con carga emotiva.

Procesamiento de Preguntas

La función de éstas es la de neutralizar lo que se considera parte de la pregunta, y la razón de hacer esto en el proceso es para mejorar la precisión en la detección de comentarios neutros (que son más difíciles de reconocer). Cualquier texto que se encuentre antes del signo de interrogación ("?") hasta un delimitador— que puede ser cualquiera de los siguientes: ",", "?", "¿", "!", ";",— se considera neutro y se modifica su valor a 0 exceptuando las palabras que tienen una carga emotiva muy negativa (-2 o -3). Por ejemplo la pregunta "¿Por qué el nuevo Nokia está tan horrible?" sigue teniendo un sentimiento negativo a pesar de ser una pregunta. Por otro lado, la pregunta, "¿Alguien conoce una excelente podadora de césped que pueda comprar?", se considera neutra. Con la incorporación de este procesamiento en el algoritmo se consigue mejorar la precisión para detectar comentarios neutros, aunque este paso necesita seguirse trabajando y se hará en trabajos futuros.

Procesamiento de Modificadores de valencia

Los modificadores de valencia, propuestos por Polanyi y Zaenen, pueden aumentar, disminuir o neutralizar la carga emotiva de una palabra [83]. Ellos suman o restan una constante c a la carga de la palabra que están modificando; utilizaron $c = 1$ en ese trabajo. Balahur y Montoyo utilizan un coeficiente $k = 1.5$ que multiplica la carga emotiva de la palabra cuando la aumenta ó $k = 0.5$ cuando la disminuyen [7]. De igual forma, Jang y Shin utilizan el coeficiente $k = 2$ cuando la carga aumenta y $k = 0.5$ cuando la carga diminuye [48]. Nosotros utilizamos estos últimos coeficientes para aumentar o disminuir la carga emotiva de una palabra. Los modificadores de valencia pueden afectar la carga emotiva de palabras anteriores o posteriores. Por ejemplo, en la frase "la película estuvo algo interesante", la palabra "algo" precede a "interesante", sin embargo, en la frase "la película me disgustó un poco", la palabra "poco" está después de "disgustó", que tiene una carga negati-

va. Un análisis sintáctico-semántico del texto podría ayudar a decidir en el caso de que existan palabras con carga emotiva justo antes y después del modificador de valencia. Nosotros utilizamos un método más simple, el cual consite en asignar un rango de "acción" a cada modificador de valencia, modificando aquellas palabras o frases que se encuentren dentro de dicho rango. Los rangos utilizados son de una palabra anterior, una palabra posterior o ambas. En futuros trabajos buscaremos implementar el análisis mencionado.

Otro modificador es el caso de la "negación", el cual cancela o invierte la carga emotiva. Pang et al. implementan la negación ajustando la carga emotiva de todas las palabras que están entre la palabra "no" y el siguiente signo de puntuación [80]. En nuestro caso implementamos la negación sólo para el caso de palabra "no" que actúa en un rango de tres palabras siguientes. El modificar "no" cancela la carga emotiva de palabras que tengan una carga $c >= -1$; aquellas que tienen una carga $c <= -2$ no son canceladas completamente, sino que se suma 1 a su valor c. Esto permite que frases como "no seas tonto", que contienen palabras muy negativas puedan seguir siendo clasificadas como negativas. Sin embargo, frases como "no es genial", son clasificadas como neutras. La implementación de la negación en análisis de sentimiento está siendo investigada actualmente [111, 21].

Los modificadores de valencia podrían tener una carga emotiva por sí mismos, como es el caso de la palabra "bien". En la frase "la película está bien padre", la palabra "bien" intensifica el valor positivo de "padre". Sin embargo, en la frase, "En el día de hoy me parece bien ir a caminar", la palabra "bien" se utiliza para expresar un sentimiento positivo. Los modificadores que tienen una carga emotiva por sí mismos y no tienen otras palabras en su rango de acción aportan su carga al sentimiento del comentario.

3.6.4 Evaluación final

La decisión final del sentimiento del comentario se toma a partir de la información que se produjo en las etapas anteriores. En estas etapas se obtuvo un peso o carga emotiva para palabras individuales, frases y preguntas. Aunque en el comentario existan más palabras con una carga emotiva, éstas pudieron ser neutralizadas.

En la evaluación se lleva a cabo una suma de los pesos de palabras y frases finales. Un criterio simple para decidir si el comentario es positivo, negativo o neutro es contar la cantidad de palabras positivas y negativas; si hay más palabras positivas que negativas el comentario será positivo, si hay más palabras negativas sera negativo y si tienen la misma cantidad será neutro. Este método se conoce como conteo de términos (*term-counting*, en inglés) [54]. La idea fue propuesta inicialmente por Turney [102, 104] aunque el método deja poco margen para los comentarios neutros. También se podría utilizar un rango $[\alpha, \beta]$ de valores que corresponderían a la categoría del neutro como lo mencionan Kennedy y Inkpen [54]. Si el valor sobrepasa a β el comentario sería positivo; y si por el contrario es menor a α, el comentario sería negativo.

Para TOM elegimos $\alpha = 0$ y $\beta = 1$ para el rango de comentarios neutros. Este rango podría ser ajustado en el futuro. Fue elegido debido a que observamos que algunos comentarios neutros del conjunto de prueba tenían palabras con un peso ligero positivo. Por ejemplo, las palabras como “comediante”, “baile”, “invitado”, “saludo” tienen un peso de 1. Pensamos que cuando existe una sola palabra de este tipo en el comentario es más probable que el comentario sea neutro.

Capítulo 4

Experimentos

En este capítulo explicamos los experimentos realizados que tienen el fin de comprobar la validez de las hipótesis planteadas. Este capítulo se divide en dos secciones principales, que son: la experimentación con TOM y la experimentación con la cuantificación del interés.

4.1 Introducción

Para llevar a cabo la experimentación con respecto a la cuantificación del Interés fue necesario crear una herramienta llamada TOM, que clasifica un comentario como positivo, negativo o neutro de acuerdo a lo descrito en la Sección 3.6; a partir de un repositorio grande de comentarios del microblog Twitter. Como veremos más adelante, se extrajeron dos conjuntos de datos: uno para extraer manualmente reglas de clasificación (a manera de un conjunto de *entrenamiento*, pero con un enfoque no supervisado) y otro para evaluar la precisión de la herramienta (a manera de *conjunto de prueba*). La precisión también fue comparada con una herramienta existente para análisis de sentimiento en español llamada Sentitext. Viendo que la precisión de TOM era aceptable, se procedió a evaluar todos los comentarios del repositorio para efectuar la experimentación relacionada con el Interés de un usuario en un tema, para la cual obtuvimos tres conjuntos de usuarios cuyas representaciones para cada grupo varían según la Tabla 4.12.

El interés de los usuarios en cada grupo fue obtenido para 38 temas y com-

parado con la probabilidad de que dos usuarios sean contactos cuando tienen un Interés alto en el mismo tema. Se conoce como **principio de homofilia** al hecho de que los contactos o interacciones son más frecuentes entre personas similares que en personas menos similares [17, 69]. Este principio ha sido estudiado en redes sociales donde se observa que personas con intereses similares tienden a unirse [108, 23]. De Choudhury encontraró que el interés en los mismos temas se asocia a un alto grado de homofilia entre los usuarios del microblog Twitter (nuestro caso de estudio) [22]. Basados en estos trabajos podemos validar los resultados de la cuantificación del Interés a través de la existencia de una correlación fuerte o moderada con la probabilidad de que los usuarios sean contactos cuando tienen un Interés muy positivo o muy negativo por un tema.

4.2 Experimentos de análisis de sentimiento

En esta sección describiremos los experimentos realizados con respecto al análisis de sentimiento que se lleva a cabo mediante la herramienta TOM que desarrollamos. Describiremos la forma en que se obtuvieron los comentarios del microblog Twitter y como se creó el conjunto de palabras (léxico) con carga emotiva (positiva o negativa) así como el conjunto de modificadores de valencia y frases. Evaluaremos la precisión de TOM con respecto a tres evaluadores humanos, quienes clasificaron manualmente dos conjuntos de comentarios. El primero de estos conjuntos fue utilizado como *conjunto de entrenamiento* para la obtención de las reglas manuales y el segundo utilizado como *conjunto de prueba* con el cual se evalúa la precisión. Finalmente, compararemos la herramienta TOM con otra herramienta existente llamada Sentitext diseñada para trabajar con texto en español. La comparación se llevó acabo midiendo la precisión de cada herramienta con el mismo *conjunto de prueba*.

4.2.1 Clasificación de sentimiento

En esta sección describiremos el proceso de creación del léxico y de la herramienta TOM. Posteriormente evaluaremos su desempeño en comparación con los tres evaluadores humanos.

Configuración

El repositorio de comentarios es un conjunto de comentarios provenientes de usuarios del microblog Twitter. Twitter ofrece un conjunto de funciones o métodos llamado API (Interfaz de Programación de Aplicaciones) que permiten la comunicación de una computadora con la red social. Una API es una interfaz de comunicación entre dos componentes de software que facilita la labor de comunicación y programación. Utilizando la API de Twitter nos fue posible crear un programa para obtener 40,186,542 comentarios provenientes de 80,954 usuarios. Los comentarios se extrajeron selectivamente eligiendo primero a los usuarios y después extrayendo 1,000 comentarios más recientes de cada uno de ellos. Si el usuario no había publicado aún 1,000 comentarios se tomaba la cantidad que tuviera.

Los usuarios que se registran en Twitter tienen un espacio entre sus datos personales que les permite publicar su lugar de origen o el lugar donde se encuentran en ese momento. Para la selección de usuarios elegimos aquéllos que habían escrito cualquiera de las palabras que aparecen en la Tabla 4.1, las cuales hacen referencia a la ciudad de Monterrey, N.L. México y su área metropolitana (note que son los nombres y abreviaciones de municipios colindantes). El proceso de extracción comienza mediante la selección de un conjunto de 100 usuarios que sirvieron como *semilla* extraídos de Twitter mediante una búsqueda de comentarios que incluyeran la palabra "Monterrey". Partiendo de estos usuarios se exploró en sus contactos en busca de aquellos que incluyeran las palabras mencionadas en la Tabla 4.1 en su lugar de origen. De cada usuario nuevo encontrado (que fue elegido por tener la mayor cantidad de contactos que mencionaban algún municipio del área metropolitana) se

Tabla 4.1: Palabras utilizadas para seleccionar usuarios.

Apodaca	Escobedo
Guadalupe	Monterrey
Monterrey, N.L.	Mty
SanNicolas	San Nicolas
San pedro	Santa Catarina

extrajeron también sus contactos y se repitió el proceso hasta obtener la muestra completa.

Twitter ofrece un segundo método para la seleccion de usuarios que es a través de la lectura de la ubicación geográfica para aquellos que cuentan con dispositivos que tienen un sistema de Posicionamiento Global (del inglés, Global Positioning System o GPS) integrado como algunos de los celulares recientes. Ésta sería la forma idónea para seleccionar usuarios de una cierta región ya que se podría detectar por medio del GPS a los usuarios que se encuentren en ella con una gran precisión. La dificultad de esto radica en que son muy pocos los usuarios que permiten a sus dispositivos compartir ese dato,— quizá por cuestiones de seguridad personal— lo que nos obliga a recurrir al método mencionado más impreciso. En la muestra inicial de 100 usuarios que obtuvimos de Twitter sólo el 14.6 % de los comentarios producidos por ellos tenían información sobre la ubicación, mientras que en la muestra global obtenida posteriormente, y que incluye a los 80,954 usuarios, el porcentaje es de 1.05 %. Pensamos que la diferencia en el porcentaje se debe a la forma en que fueron elegidos los primeros 100 usuarios y al funcionamiento de Twitter el cual ofrece resultados según su algoritmo de búsqueda interno.

A partir del repositorio de comentarios se eligieron al azar dos conjuntos de 3,000 comentarios cada uno que cumplían con la condición de ser escritos por un usuario diferente para evitar seleccionar comentarios escritos de forma continua en una conversación personal, ya que este tipo de conversaciones son comunes en Twitter. Llamaremos a estos conjuntos C_1 y C_2 para hacer referencia a ellos en este

capítulo. Cada uno de los comentarios en ambos conjuntos fue clasificado manualmente utilizando una herramienta de tecnologías de la información creada por el autor de esta tesis que trabaja a través de Internet. Fueron tres personas quienes clasificaron ambos conjuntos de 3,000 comentarios C_1 y C_2. Dos de las personas que evaluaron los comentarios son estudiantes de Maestría (incluído el autor de esta tesis) y una Doctora (asesora de esta tesis). Los evaluadores no tuvieron comunicación entre ellos durante el tiempo en que clasificaron los comentarios y no se definió ningún tipo de criterio para decidir cuando un comentario sería positivo, negativo o neutro. Esto permitió que los evaluadores pudieran clasificar libremente, sin ningun tipo de prejuicio en cuanto al sentimiento. Definir un criterio de clasificación podría ser riesgoso en el sentido de que los evaluadores estarían utilizando mentalmente un método de cálculo de proximidad hacia dicho criterio, lo que finalmente producirá un sesgo en los resultados si no se define apropiadamente el criterio. La Figura 4.1 muestra una imagen de la herramienta utilizada por los evaluadores durante la clasificación manual de los comentarios.

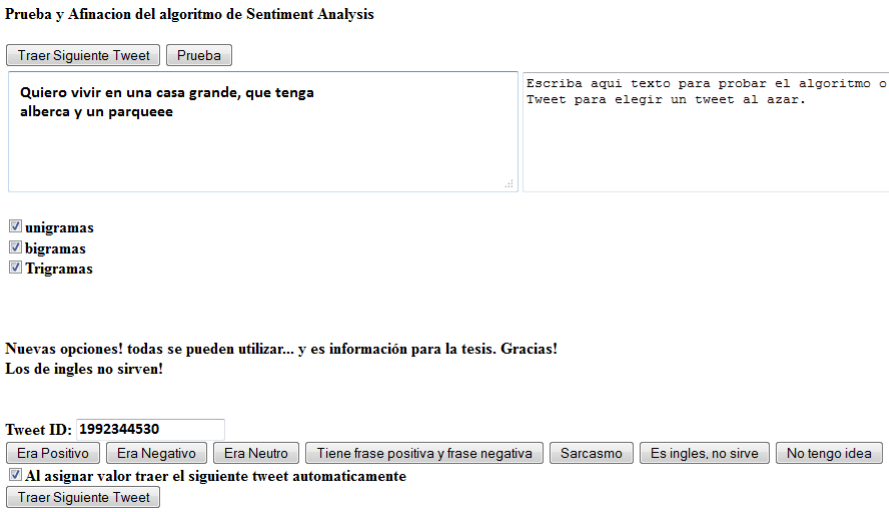


Figura 4.1: Imagen de la herramienta de clasificación manual utilizada por los evaluadores.

La herramienta permitía clasificar los comentarios de diferentes formas como se muestra en la Tabla 4.2. Se optó por asignar como categoría definitiva a un comentario aquella que hubiera sido asignada por los tres evaluadores de forma unánime

desechando aquellos comentarios en los que no se estuvo de acuerdo. Además de las clasificaciones obvias (positiva, negativa y neutra) consideramos oportuno identificar comentarios que presentaban dos opiniones al mismo tiempo, una negativa y una positiva, así como identificar comentarios sarcásticos y aquellos que el evaluador no fue capaz de decidir la categoría a asignar considerándolos difícil de clasificar.

Tabla 4.2: Opciones disponibles para la clasificación manual.

Evaluación	Evaluador
Positivo	Tuvo una sensación positiva respecto al comentario
Negativo	Tuvo una sensación negativa respecto al comentario
Neutro	No considera que el comentario tenga un sentimiento
Positivo y Negativo	Cree que parte del comentario es positiva y parte es negativa
Sarcasmo	Cree que el autor del comentario está siendo sarcástico
Otro idioma	Detectó un comentario en otro idioma
No sé	No ha podido decidir si el comentario tiene un sentimiento o no

Se utilizó el conjunto C_1 para generar una lista de palabras, frases y modificadores de valencia que se introducirían al diccionario de TOM (conjunto de reglas manuales). El conjunto C_2 fue apartado para la experimentación.

El primer experimento consistió en utilizar TOM para clasificar automáticamente los comentarios del conjunto C_2, los cuales no se han utilizado para desarrollar el algoritmo. Se tomaron en cuenta los comentarios que fueron clasificados de la misma forma por los tres evaluadores para medir la precisión de TOM al clasificar el sentimiento. La precisión fue obtenida dividiendo los aciertos (cuando TOM clasifica igual que los evaluadores) por el total de estos comentarios. Suponemos que cualquier persona evaluaría estos comentarios de dicha manera y por consiguiente un algoritmo de clasificación automática debería hacerlo también. Jiang et al. utilizaron un conjunto de comentarios clasificados manualmente como positivos, negativos o neutros por dos evaluadores humanos para validar su método de análisis de sentimiento hacia un objetivo (o *target*) [52]. Otros comentarios que los evaluadores consideraron difíciles de clasificar como: sarcasmo, positivo y negativo, también fue-

ron excluídos al evaluar la precisión centrandonos en las categorías de sentimiento positivo, negativo y neutro.

En la experimentación con TOM incluímos un análisis de la contribución de cada una de las partes del algoritmo a la precisión. Esto es, analizamos los cambios en la precisión al activar o desactivar funciones de TOM tales como el léxico de Liu, léxico de Twitter, carga emotiva de las palabras (pesos), frases, modificadores de valencia, corrector de ortografía, procesamiento de preguntas y eliminación de sustantivos.

Al desactivar los pesos en el léxico (rango de -3 a +3), se utiliza un rango simple de -1 a +1 para palabras negativas (-1), neutras (0) y positivas (+1). Es decir, palabras que son muy positivas o muy negativas tendrán el mismo peso que una palabra con carga emotiva más ligera. Por ejemplo, las palabras "Perfecto" (peso = +3) y "tranquilo" (peso = +1) tendrán el mismo peso de +1.

La experimentación con respecto a los sustantivos se debe a que observamos en el conjunto de comentarios C_1 que algunos sutantivos pueden tener carga emotiva, pero que ésta se vuelve neutra al ser utilizados de cierta forma en un comentario; por ejemplo, en la frase "su familia se encuentra en casa", la palabra "familia" no tiene sentimiento. Sin embargo, en la frase "juntos son como una familia", la palabra "familia" tiene un sentimiento positivo. El análisis de discurso es un procesamiento realizado en un texto el cual asigna etiquetas a cada una de las palabras según su función dentro del comentario [15]. Mediante análisis de discurso podemos distinguir los sustantivos, verbos, artículos, adjetivos, pronombres, etc. lo cual brinda información que puede ser utilizada en un procesamiento posterior (como análisis de sentimiento). Nosotros no hacemos un análisis de discurso completo, pero utilizamos una lista de palabras que neutralizan la carga emotiva de la siguiente palabra. Las palabras que utilizamos con este fin, son: "la", "'el", "sin", "ni", "mi", "su", "tu", "mis". La siguiente palabra es neutralizada siempre y cuando su carga emotiva sea $>= -1$. Por ejemplo, las expresiones "la traición", "el amor", "tu caída", "ni ayuda ni perjudica", serían clasificadas como neutras.

En la siguiente sección presentamos los resultados obtenidos.

Resultados

En la Tabla 4.3 se muestran la cantidad de comentarios en los que dos y tres evaluadores coinciden en la clasificación y el porcentaje en relación al total de comentarios tomando en cuenta que por lo menos un evaluador lo clasificó como se indica en la tabla para cada caso. Es decir, si un evaluador determina un cierto sentimiento para un comentario, se calcula la probabilidad de que otro también lo haga.

Para el conjunto C_1, teniendo en cuenta solo sentimientos positivos, negativos y neutros, obtuvimos que en promedio dos evaluadores están de acuerdo en la clasificación en el 57.72 % de los casos cuando uno de ellos ya lo clasificó de dicha manera, y los tres evaluadores estarán de acuerdo en un 24.64 % de los casos. Para el conjunto C_2 obtuvimos que dos evaluadores estarán de acuerdo en el 59.18 % de los casos y tres evaluadores estarán de acuerdo en 28.86 % de los casos.

La precisión de TOM fue medida utilizando solamente el conjunto C_2 ya que éste no fue utilizado durante su creación. La Tabla 4.4 muestra los porcentajes para el sentimiento positivo, negativo y neutro. Obtuvimos que la precisión global de TOM es de 77.32 %. En este conjunto encontramos 1,147 casos en los que los tres evaluadores coinciden en cuanto al sentimiento, los cuales representan un 38.23 % de la muestra y fueron utilizados para medir la precisión. El 11.60 % de los comentarios fueron clasificados como otro idioma y en el 50.17 % restante los evaluadores no coincidieron en la clasificación. Es interesante observar que los evaluadores no coinciden en los casos de positivo y negativo, sarcasmo y no sé, por lo que no podemos afirmar que algún comentario de nuestra muestra pertenece a estas categorías.

La Tabla 4.5 muestra el detalle de los errores en la clasificación de TOM. Observamos que cuando un comentario es positivo (o negativo), es menos probable que se clasifique como negativo (o positivo) en comparación con los comentarios

Tabla 4.3: Coincidencia en la clasificación de los evaluadores humanos.

Conjunto	Sentimiento	Uno	Dos	%	Tres	%
	Positivo	526	465	55.39 %	188	15.95 %
	Negativo	398	331	66.55 %	461	38.74 %
	Neutro	898	589	51.22 %	354	19.23 %
C_1	Positivo y Negativo	83	18	17.82 %	0	0.00 %
	Sarcasmo	36	3	7.69 %	0	0.00 %
	No sé	92	2	2.13 %	0	0.00 %
	Otro idioma	42	61	90.21 %	326	75.99 %
	Positivo	406	320	57.08 %	220	23.26 %
	Negativo	362	223	55.53 %	229	28.13 %
	Neutro	696	590	64.92 %	698	35.18 %
C_2	Positivo y Negativo	54	12	18.18 %	0	0.00 %
	Sarcasmo	81	8	8.99 %	0	0.00 %
	No sé	355	11	3.01 %	0	0.00 %
	Otro idioma	97	68	81.09 %	348	67.84 %

Tabla 4.4: Precisión de TOM.

Sentimiento	Tres	TOM	Precisión
Positivo	220	171	77.73 %
Negativo	229	159	69.43 %
Neutro	698	592	84.81 %
Global			77.32 %

Tabla 4.5: Errores de clasificación de TOM.

			Incorrectos					
Clase	Comentarios	Correctos	Positivos		Negativos		Neutros	
Positivos	220	171	0	0.00 %	6	2.72 %	43	19.54 %
Negativos	229	159	11	4.80 %	0	0.00 %	59	25.76 %
Neutros	698	592	46	6.59 %	60	8.60 %	0	0.00 %

neutros.

Para saber en qué beneficia cada una de las partes del algoritmo (léxico de Liu, léxico de Twitter, carga emotiva de las palabras (pesos), frases, modificadores de valencia, corrector de ortografía, procesamiento de preguntas y eliminación de sustantivos) hemos desactivado cada una de las funciones por separado al momento de hacer la evaluación de los comentarios. Los resultados se ven en la Tabla 4.6.

Discusión de los resultados

El análisis de sentimiento es una tarea compleja ya que tiene que ver con la opinión. El porcentaje de coincidencia de los evaluadores con respecto al sentimiento del comentario es bajo, y conforme más personas se unen a la evaluación observamos una tendencia a disminuir aún más. A partir de los resultados podemos afirmar que aún para evaluadores humanos es difícil ponerse de acuerdo si un comentario debe considerarse positivo, negativo o neutro. Este hecho tiene implicaciones en la investigación, sobre todo a la hora de definir un procedimiento o criterio para medir la precisión con la que se evalúa un comentario automáticamente; además es evidencia de que la evaluación se lleva a cabo bajo el contexto del conocimiento y/o experiencia personal del evaluador. Incluso para la clasificación de comentarios en otros idiomas existen diferencias entre un evaluador y otro. En este caso la clasificación es mucho mas objetiva puesto que se trata de identificar un idioma, sin embargo para el conjunto C_2 la coincidencia de tres evaluadores es de un 67.84 %. Creemos que se debe a que no se definió un criterio para la clasificación de comentarios que

Tabla 4.6: Precisión de TOM al activar/desactiar funcionalidades.

Clase	Solo Liu	Solo Twitter	Liu + Twitter
Positivos	25.91 %	45.91 %	53.18 %
Negativos	50.22 %	58.95 %	58.95 %
Neutros	86.82 %	88.68 %	85.82 %
Global	54.32 %	64.51 %	65.98 %
Clase	Con Pesos	Ortografía	Modificadores
Positivos	72.27 %	74.09 %	75.91 %
Negativos	67.69 %	69.00 %	69.87 %
Neutros	83.95 %	83.95 %	83.67 %
Global	74.64 %	75.68 %	76.48 %
Clase	Frases	Sin Preguntas	Sin Sustantivos
Positivos	77.73 %	77.73 %	73.18 %
Negativos	70.31 %	69.43 %	70.31 %
Neutros	83.38 %	84.81 %	85.67 %
Global	77.14 %	77.32 %	76.39 %

mezclan el idioma inglés y español ya que existen frases en inglés que aparecen con frecuencia en los comentarios, como: "good morning", "hello", "perfect!", etc. En los casos de sarcasmo, comentarios con sentimiento positivo y negativo a la vez, y comentarios "difíciles" no hubo coincidencia por los tres evaluadores. Esto quiere decir que son más difíciles de detectar y también será más difícil obtener un conjunto suficientemente grande para su análisis.

La precisión global de TOM fue de 77.32 %. Este resultado es aceptable, sin embargo, creemos que la precisión puede mejorar si se continua expandiendo su diccionario. De la Tabla 4.6 se puede observar que los mecanismos que principalmente contribuyen a la precisión son:

- Las 2318 palabras iniciales provenientes de Twitter.
- Los pesos asignados a cada palabra del diccionario.

Esto sugiere que el conjunto de palabras de TOM podría ser ampliado con el fin de mejorar la precisión. Utilizando solamente el diccionario de Liu tenemos una precisión de 54.32 %, y utilizando solo el léxico de Twitter obtenemos una precisión de 64.51 %. Esto quiere decir que para trabajar con microblogs donde las personas pueden escribir en la forma que lo deseen, los errores de ortografía o gramática son muy comunes y existen expresiones coloquiales (como "amors", "chido", etc.), es útil contar con un diccionario que incluya dichos términos. El léxico de Liu contribuye en 1.47 % a la precisión cuando se combina con el léxico de Twitter que obtuvimos alcanzando un 65.98 %.

La asignación de pesos a las palabras en el rango de -3 a +3, proporciona otro incremento significativo en la precisión llegando a 74.64 %. Esto quiere decir la asignación de pesos mayores a palabras con carga emotiva mayor contribuye en gran medida en la precisión.

La corrección de ortografía, modificadores de valencia, procesamiento de frases y procesamiento de preguntas contribuyen un 2.68 % a la precisión con respecto al uso

del léxico de Liu + Twitter y pesos de las palabras. Creemos que estas características forman parte de un conjunto de procesamientos más finos cuyo perfeccionamiento será necesario para que en un futuro TOM pueda tener niveles de precisión más altos. Sin embargo, la precisión alcanzada solo por estas tres características es buena (74.64 %).

La eliminación de los sustantivos no contribuye a mejorar el resultado global. Al excluir los sustantivos obtenemos una pérdida en la precisión global de 0.93 %. Este resultado es congruente con los trabajos de Pang et al. [80] y Go et al. [3], ya que el análisis del discurso (part-of-speech) que realizaron en los comentarios no contribuyó a un mejor resultado. Sin embargo Kouloumpis et al. mencionan que es necesario seguir investigando su utilización en análisis de sentimiento [57]. Con lo cual, nosotros estamos de acuerdo.

Algunas de las complicaciones con las que nos encontramos en la evaluación son por ejemplo, el sarcasmo, la detección de chistes o bromas, dichos comunes, y principalmente cuando se necesita un conocimiento previo de lo que se está hablando. Las bromas son difíciles de detectar y de clasificar; la clasificación depende del tipo de broma. Por ejemplo, si se hace una broma acerca de una persona quizá esto represente una opinión negativa o positiva hacia dicha persona. Si se trata de un chiste podría ser neutro aunque en la composición del chiste se utilicen palabras emotivas. El caso del sarcasmo ha sido mencionado por algunos autores y actualmente se encuentra en desarrollo la detección y clasificación de este tipo de comentarios [38]. Estos comentarios necesitan que el lector tenga un conocimiento de lo que se habla para poder detectar el sentimiento detrás de las palabras; por ejemplo la frase: “Hay una fila que parece que hoy dan caviar y langosta...”. En esta frase se debe saber de antemano que a la mayoría de las personas les disgusta hacer fila para ser atendidos y también que el caviar y la langosta son bien apreciados como alimentos. Mediante un análisis que incluya dicha información podríamos concluir que en el comentario existe una queja de la fila expresándose en forma de sarcasmo; en este caso es un sentimiento negativo.

Dado que TOM es una herramienta de análisis de sentimiento, ésta podría ser utilizada en otras aplicaciones distintas a la cuantificación del interés; tales como: análisis de mercados, detección de opiniones sobre un producto, pronósticos con relación a la bolsa de valores con base en el sentimiento de los usuarios en microblogs [11], o predecir las ventas de una película [73].

En la siguiente sección compararemos el desempeño de TOM con una herramienta para análisis de sentimiento en español llamada Sentitext.

4.2.2 Experimentos comparativos

En esta sección compararemos a TOM con una herramienta existente llamada Sentitext [76] que evalúa el sentimiento de texto en español. Sentitext es una de las pocas herramientas que existen para determinar el sentimiento de un texto en español. Su funcionamiento está basado en un diccionario de palabras, al igual que TOM, y en un conjunto de reglas del lenguaje. Utilizaremos el conjunto de comentarios C_2 para evaluar la precisión de Sentitext con respecto a los tres evaluadores humanos y compararemos dicha precisión con la de TOM.

Configuración

Sentitext asigna un número de "estrellas" en vez de una polaridad o neutralidad del sentimiento como lo hace TOM. Por esta razón dividimos el rango de "estrellas" y los asociamos a un tipo de sentimiento. Sentitext arroja valores entre 0 y 10 estrellas, donde 0 quiere decir que es un comentario muy negativo y 10 que es un comentario muy positivo. Si la evaluación obtenida con Sentitext era de 5, consideramos el comentario como neutro dado que Sentitext no fue capaz de encontrar palabras con carga emotiva. De igual forma si el resultado era menor o mayor de 5 lo consideramos como negativo o positivo respectivamente. El criterio para elegir la clasificación se ve en la Tabla 4.7. Una descripción más completa de Sentitext puede encontrarse en [77].

Tabla 4.7: Criterio de conversión de estrellas a polaridad de sentimiento.

Estrellas	Sentimiento
<5	Negativo
0	Neutro
>5	Positivo

Tabla 4.8: Comparación de evaluación de TOM con Sentitext.

Sentimiento	TOM	Sentitext	Diferencia
Positivo	77.40 %	75.48 %	1.92 %
Negativo	68.33 %	62.44 %	5.89 %
Neutro	85.01 %	78.46 %	6.55 %
Global	76.91 %	72.13 %	4.79 %

Sentitext presentó un error al evaluar ciertos comentarios arrojando un valor nulo— quizá debido al uso de caracteres especiales en los comentarios tales como $, #, &, , etc. Dichos casos en los que sentitext arrojó un error fueron excluídos al evaluar a Sentitext. En total fueron 31 casos con error que se excluyeron de la comparación. La precisión de TOM fue recalculada después de eliminar estos casos.

Resultados

Para el conjunto C_1 obtuvimos que Sentitext coincide con los tres evaluadores en promedio en el 73.88 % de los casos y para el conjunto C_2 obtuvimos que coincide en promedio en el 72.13 % de los casos. Utilizando nuestro evaluador TOM el cual contiene palabras específicas que se utilizan en microblogs, obtuvimos que para el conjunto C_2 se clasificaron correctamente en promedio el 76.91 % de los comentarios. Esto representa una mejora de 4.79 %. En la Tabla 4.8 se ve una comparación detallada de las clasificaciones obtenidas por TOM y por Sentitext.

En las Figuras 4.2, 4.3 y 4.4 se ilustra el comportamiento de ambos algoritmos (TOM y Sentitext) durante la evaluación de los comentarios; se observa como varia

la precisión de la clasificación. Mostramos solo las gráficas para el conjunto C_2. Las mediciones en la gráfica corresponden a intervalos de 20 comentarios hasta completar el total de comentarios en los que los tres evaluadores coinciden.

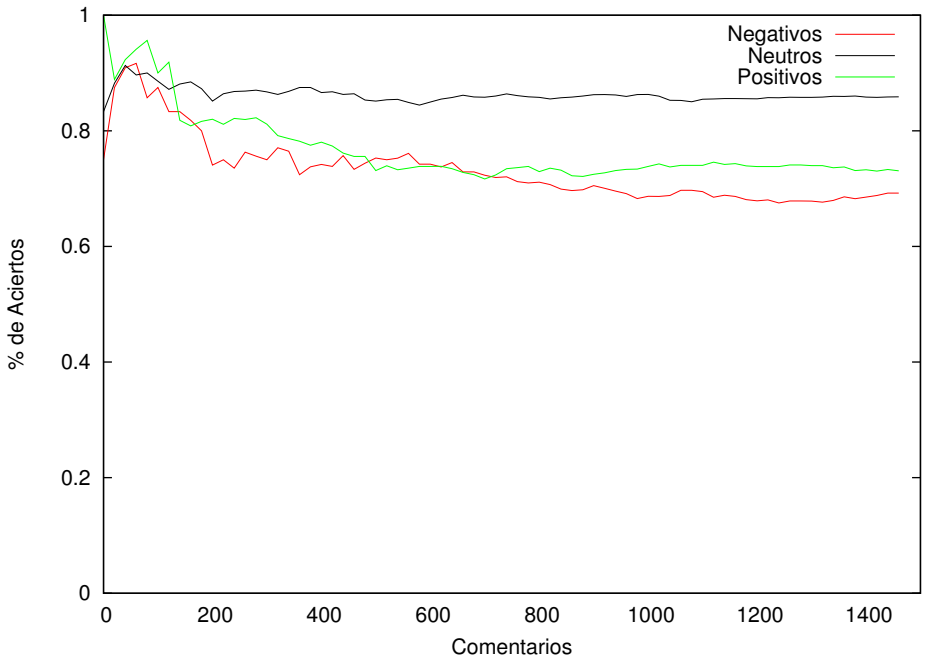


Figura 4.2: Precisión de TOM con respecto al número de comentarios evaluados.

El mecanismo de corrección de ortografía de TOM le permite encontrar palabras que no se encuentran en el diccionario tal cual fueron escritas. Por ejemplo, en el comentario "cada vez me gusta menos FB... aburridoo" se observa la palabra "aburridoo" que fue detectada correctamente por TOM. En la Tabla 4.9 se ve una lista de comentarios que fueron clasificados correctamente por TOM e incorrectamente por Sentitext. En algunos casos TOM toma ventaja al identificar correctamente estas palabras. Note también el comentario "Buen dia que Dios los bendiga", el cual no refleja una opinión sobre algo (como un producto) pero manifiesta un buen deseo y los evaluadores estuvieron de acuerdo en que es un comentario positivo. Sin embargo, Sentitext no encuentra palabras de opinión y por tanto lo clasifica como neutro. Diferencias como ésta pueden deberse a la existencia de acuerdos acerca de lo que es una opinión o una sensación positiva o negativa; en nuestro caso utilizamos la definición propuesta por Go et al. [3].

Tabla 4.9: Comparación de TOM (correcto) con Sentitext (incorrecto).

Comentario	Evaluadores	TOM	Sentitext
RT @criistto: Pido una oracion en cadena por mi amigo @davidalvareez que mañana presenta, RT porfa :P	Neutro	Neutro	Positivo
Felicidades don Oscar #LaVozMéxico.	Positivo	Positivo	Neutro
Por cierto, hoy me ire a dormir con mi nuevo mantra: Me importa un culo hago lo que sea, me importa un culo hago lo que sea. Gud nait	Negativo	Negativo	Neutro
RT @A_reirse: El cura en la iglesia dice: -Hoy confesare a las devotas. Se levanta una mujer y pregunta: -padre, ¿y las que vinimos en s ...	Neutro	Neutro	Positivo
@simonnin jajajaja de seguro fue una de tus primeras fotos	Neutro	Neutro	Positivo
cada vez me gusta menos FB... aburridoo	Negativo	Negativo	Positivo
En tiempos de paz lo hijos entierran a los padres, en tiempos de guerra los padres entierran a los hijos. Pelicula de los INMORTALES .	Neutro	Neutro	Negativo
?@PPmerino: Dato curioso: Han inmigrado a México 452 ciudadanos de Corea del Norte desde 1995.? wuoraleee son un buen!!	Neutro	Neutro	Positivo
?@malatorre: MITAD Y MITAD LUCHA POR EL AMOR.....YA VIENE...SI EL PRIMERO ARRASO EL SEGUNDO APLASTARA a q mugrero ya cambiale no jodas	Negativo	Negativo	Positivo
Buen dia que Dios los bendiga	Positivo	Positivo	Neutro

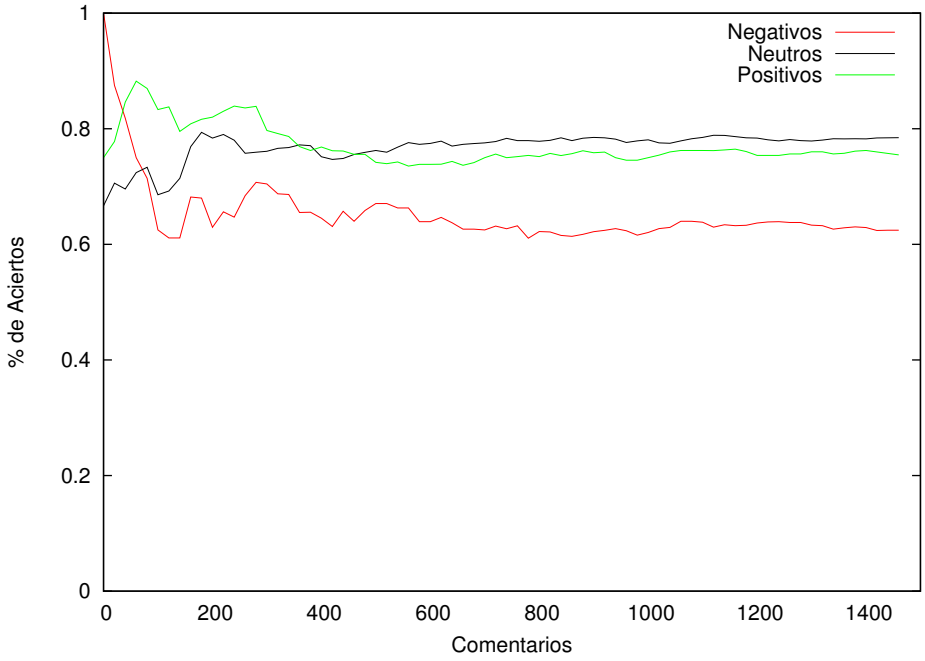


Figura 4.3: Precisión de Sentitext con respecto al número de comentarios evaluados.

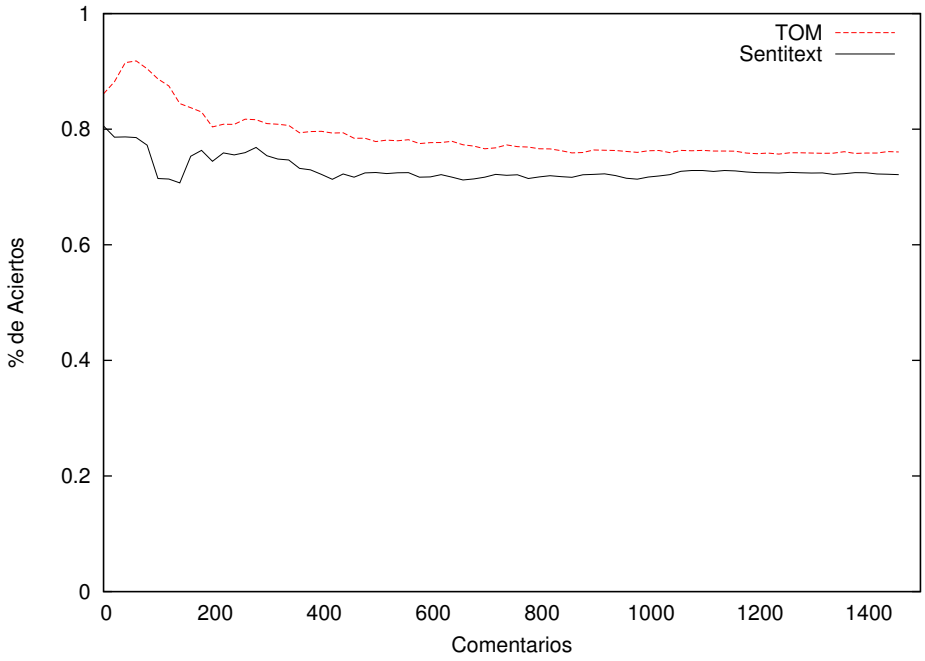


Figura 4.4: Precisión promedio de Sentitext y TOM.

Así mismo la Tabla 4.10 muestra un conjunto de comentario donde Sentitext hizo la clasificación correcta. En estos ejemplos notamos que algunas palabras como

"pachorruda" o "reirse" hacen falta en el diccionario de TOM, y algunas formas del lenguaje como en la expresión "Feliz como una lombriz" no son bien identificadas.

Discusión de los resultados

El desempeño de TOM es comparable al de Sentitext, teniendo una ventaja de 4.79 % en la precisión global al trabajar con comentarios cortos publicados en microblogs como Twitter. Esto quiere decir que la herramienta puede ser utilizada en trabajos de minería de texto en microblogs ya que cuenta con una precisión comparable a las herramientas existentes en el estado del arte que trabajan con el español. Sentitext no está diseñado para trabajar con Twitter sino que tiene un objetivo más general, enfocado a trabajar con textos más grandes y con menos deficiencias en gramática y ortografía. Esto comprueba que al trabajar con microblogs es necesaria la utilización de un léxico que incluya términos que se utilizan en ellos, aunque no formen parte del idioma con el que se trabaja propiamente dicho.

TOM clasifica correctamente 85.01 % de los comentarios neutros en comparación con 78.46 % de Sentitext. Esto es importante dado que al clasificarse una mayor cantidad de comentarios neutros correctamente se reduce la introducción de ruido a las otras categorías (positivos y negativos) lo que se traduce en una mejor extracción de información.

En la Tabla 4.9 observamos como el corrector de ortografía ayudó en algunos casos, como en la clasificación del comentario "cada vez me gusta menos FB... aburridoo"; en el cual se detectó correctamente la palabra "aburridoo". Por ahora nuestro corrector de ortografía se limita a este tipo de casos como se explicó en la Sección 3.6.3 pero otras correcciones podrían ser útiles como es el caso de los acentos o la sustitución de ciertos caracteres en la palabra. Por ejemplo, la palabra "canción", en microblogs es frecuente encontrarla escrita como "kancion" o "kanción". Mediante observación, notamos en el conjunto de entrenamiento C_1 que algunas palabras pueden corregirse sustituyendo letras que tengan un sonido parecido. Por ejemplo,

Tabla 4.10: Comparación de TOM (incorrecto) con Sentitext (correcto).

Comentario	Evaluadores	TOM	Sentitext
Por que no eres mas inteligente?	Negativo	Neutro	Negativo
Feliz como una lombriz	Positivo	Neutro	Positivo
La gente esta semana ya es totalmente pachorruda	Negativo	Neutro	Negativo
Hay que reirse un rato http://t.co/arzPHWP0	Positivo	Neutro	Positivo
http://t.co/qQVwGmWE todo. El. dia. con. Esta. cancion. Grcias a mi. Sobrina Vale	Neutro	Negativo	Neutro
De esas veces que todo se va al carajo en un segundo	Negativo	Neutro	Negativo
Esta mas que claro, Randall-El debe de entrar, Roethlisberger no esta listo, ya tiene 2 intercepciones, se nota su desesperacion...	Negativo	Neutro	Negativo
domingo 1 de enero solo drink arenoso 7:00pm	Neutro	Negativo	Neutro
Cada vez queda menos para sortear la Tablet, así que si aún no participas es el momento de hacerlo! Ingresa a... http://t.co/PiMSZacJ	Neutro	Negativo	Neutro
La diferencia entre los que viven por fe y los que no es que los primeros no permiten que sus dudas decidan.	Neutro	Positivo	Neutro

Tabla 4.11: Comparación de errores de TOM con Sentitext.

TOM

			Incorrectos					
Clase	Total	Correctos	Positivos		Negativos		Neutros	
Positivos	208	161	0	0.00 %	6	2.88 %	41	19.71 %
Negativos	221	151	11	4.98 %	0	0.00 %	59	26.70 %
Neutros	687	584	44	6.40 %	59	8.59 %	0	0.00 %

Sentitext

			Incorrectos					
Clase	Total	Correctos	Positivos		Negativos		Neutros	
Positivos	208	157	0	0.00 %	7	4.46 %	44	28.03 %
Negativos	221	138	24	17.39 %	0	0.00 %	59	42.75 %
Neutros	687	539	97	18.00 %	51	9.46 %	0	0.00 %

cambiando la “s” por la “c”, la “k” por la “c”, etc. Así mismo podrían intercambiarse los acentos en las vocales. El objetivo de esto es localizar dicha palabra en el léxico que contiene el peso (carga emotiva) para la palabra. Este enfoque podría ser investigado en un trabajo futuro.

En la Tabla 4.10 notamos que algunas palabras hacen falta en el léxico de TOM tales como: “pachorruda”, “reirse” y “carajo”. Estas palabras no se encontraron en el léxico y por lo tanto fueron clasificadas como neutras. Creemos que el léxico debe seguir ampliandose con el fin de mejorar la precisión. La palabra “Vale” en el comentario “El. dia. con. Esta. cancion. Grcias a mi. Sobrina Vale” fue clasificada como negativa, cuando en verdad podría ser el nombre de una persona o una expresión común en España. Dicha palabra sí sería negativa en un expresion como “Me vale”. Esto hace evidente que se necesita un procesamiento de desambigüación de las palabras que por ahora no está incluído en el algoritmo de TOM.

En el comentario “Esta mas que claro, Randall-El debe de entrar, Roethlisberger no esta listo, ya tiene 2 intercepciones, se nota su desesperacion...”, la palabra

"claro" fue clasificada como positiva, además de estar precedida por un modificador de valencia "que", el cual aumenta el valor multiplicandolo por 2, dando como resultado que la expresión "que claro" tenga mayor peso a la palabra negativa "desesperacion" y por lo tanto fue clasificado incorrectamente. Este caso podría solucionarse agregando la frase "más que claro" al conjunto de frases neutras del diccionario de TOM. Sin embargo, debe analizarse si esta frase es neutra en todos los casos. De igual forma en el comentario "Cada vez queda menos para sortear la Tablet", la frase "queda menos para" también podría ser agregada al conjunto de frases neutras.

En estos casos Sentitext toma ventaja sobre TOM debido a que cuenta con un léxico y un conjunto de frases y reglas del lenguage más amplios. La precisión de TOM puede mejorarse de varias formas, algunas pueden ser: ampliar el léxico, agregar más frases con carga emotiva, procesar mejor los modificadores de valencia, utilizar información de un analizador de discurso, utilizar otras alternativas al procesamiento de preguntas, incluso podría adaptarse a un enfoque de aprendizaje supervizado para la detección automática de nuevas palabras y su carga emotiva.

Los resultados obtenidos con TOM son alentadores en esta primera etapa, pensamos que en un futuro podrá ser una excelente herramienta para minería de texto en español en microblogs.

4.3 Experimentos de cuantificación del Interés

En esta sección describiremos los experimentos realizados para cuantificar el interés y explicaremos los resultados obtenidos.

4.3.1 Configuración

Para la representación de los usuarios se utilizaron comentarios provenientes de sus contactos. No hemos podido encontrar investigaciones acerca del número adecuado de contactos que debe ser considerado para obtener una representación

del usuario ni del número de comentarios a tomar de cada contacto. Por esta razón utilizamos las tres combinaciones de la Tabla 4.12:

Estrategias para la representación del usuario
20 Contactos del usuario seleccionados al azar
40 Contactos del usuario seleccionados al azar
30 Contactos que se hayan mencionado mutuamente

Tabla 4.12: Representaciones de usuario utilizadas.

Creamos tres grupos de usuarios utilizando cada una de las estrategias descritas en la Tabla 4.12 y mostramos que la probabilidad de que dos usuarios sean contactos aumenta cuando tienen un Interés alto por el mismo tema ya sea positivo (agrado) o negativo (desagrado).

Para cada estrategia obtenemos cuatro conjuntos de comentarios según el sentimiento (positivo, negativo, neutro o todos) que serán utilizados para crear el *vector de usuario* $\vec{U}$ necesario para comparar con el tema. Es decir, la representación del usuario puede crearse a partir de aquellos comentarios de sus contactos que pertenezcan solo a una clase de sentimiento. Analizaremos el efecto que tiene en el Interés el hecho de utilizar sólo un cierto tipo de comentarios ya sea positivos, negativos o neutros para representar al usuario. La Figura 4.5 muestra el modelo de experimentación utilizado.

El primer paso es calcular el Interés de cada uno de los usuarios con respecto al tema y después calcular la probabilidad de que dos usuarios sean contactos dado el nivel de Interés que tienen. Hacemos esto para los tres grupos de usuarios y para los cuatro subconjuntos de comentarios según su sentimiento.

En el caso de Twitter, además de la posibilidad de que dos usuarios se sigan el uno al otro, un usuario puede seguir a otro sin que éste último le corresponda de la misma manera. Por esta razón realizamos experimentos midiendo la probabilidad para cada clase de relación (unidireccional o bidireccional). Los vectores de palabras y valores TF-IDF fueron obtenidos utilizando la herramienta Lucene.

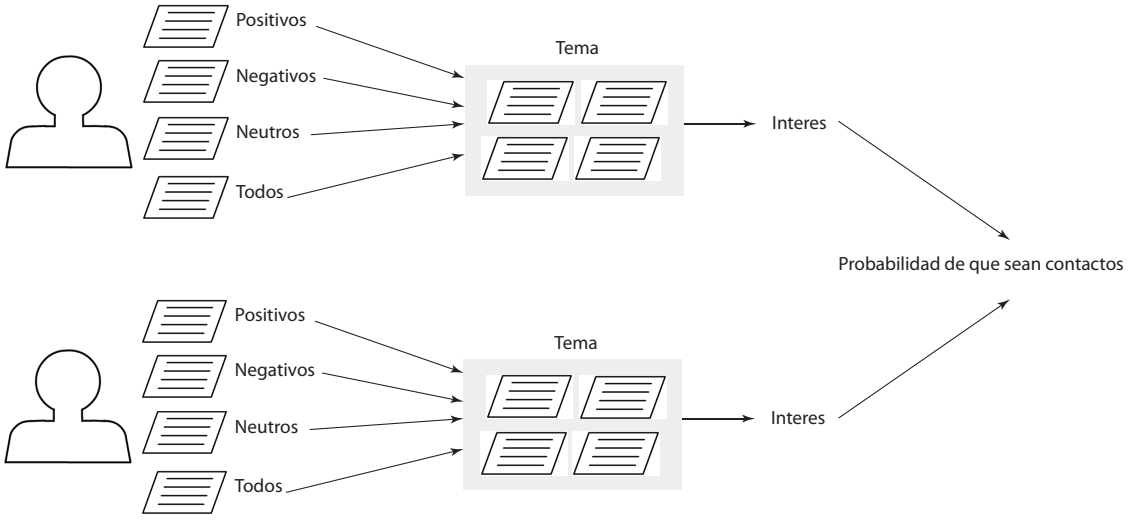


Figura 4.5: El modelo de experimentación.

Herramienta Lucene

El procesamiento se llevó a cabo utilizando la herramienta Lucene[1], la cual permite indexar documentos de texto y ha sido utilizada en diversas investigaciones que tienen que ver con la recuperación de información. En su diseño se incluye un módulo para trabajar con el idioma español necesario para los experimentos de esta tesis. Utilizando Lucene obenemos un vector de palabras que es generado automáticamente. Lucene incorpora mecanismos propios para la limpieza de texto el cual incluye una lista predefinida de *stopwords* (palabras vacías) y un mecanismo de tokenización basado en delimitadores.

Elección de Temas

Elegimos 38 temas para calcular el interés de los usuarios en el microblog Twitter. Obtuvimos 30 de ellos a partir de la lista de marcas más valiosas del mundo según la empresa Millward Brown[2] dando preferencia a marcas mexicanas. Los siguientes 8 temas fueron elegidos por ser comunes en páginas de entretenimiento

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

[2]http://www.millwardbrown.com/BrandZ/Top_100_Global_Brands.aspx

como youtube.com o yahoo.com. La Tabla 4.13 muestra los temas utilizados en los experimentos. La experimentación se realizó utilizando cada tema individual y en forma global para los 38 temas.

Marcas comerciales como temas.					
Soriana	Telmex	Microsoft	Corona	Movistar	Sanborns
CocaCola	Amazon	Cemex	Modelo	Disney	TV Azteca
Telcel	Intel	Google	Walmart	Apple	Facebook
Oracle	Toyota	Inbursa	Liverpool	IBM	Elektra
Tecate	Televisa	Bimbo	Nike	Aurrera	HP

Tabla 4.13: Estas marcas fueron elegidas por estar dentro del top 100 de marcas más valiosas según la empresa Millward Brown. Dando preferencia a las marcas mexicanas.

Temas generales	
Ciencia	Música
Cultura	Cine
Religión	Teatro
Deporte	Tecnología

Tabla 4.14: Temas generales utilizados en la experimentación. Fueron elegidos por estar presentes en algunas páginas de entretenimiento que tienen gran cantidad de visitantes como youtube.com o yahoo.com.

El *vector de tema* $\vec{T}$ fue creado haciendo búsquedas de comentarios en el repositorio completo que contuvieran algunas palabras clave dependiendo de cada tema. Estas palabras fueron obtenidas a partir del nombre del tema mediante modificaciones a su ortografía, es decir, agregando o quitando acentos y utilizando su forma singular o plural. Los temas y la cantidad de comentarios obtenidos se muestran en las Tablas 4.15 y 4.16

Marcas comerciales como temas			
Tema	Cantidad	Tema	Cantidad
Facebook	99214	TV Azteca	4418
Sorianta	4254	Movistar	1953
Google	36506	HP	4035
Amazon	3860	Televisa	20162
Apple	18951	Liverpool	3429
Telcel	15006	Walmart	2949
Cemex	1994	Aurrera	408
Tecate	1976	IBM	1465
Corona	8023	Bimbo	1332
Nike	6658	Toyota	1284
Modelo	6280	Oracle	1221
Disney	6280	Sanborns	1176
Microsoft	6172	Intel	1028
Telmex	4820	Elektra	410
Coca-Cola	4606	Inbursa	121

Tabla 4.15: Palabras utilizadas para obtener los comentarios relacionados con cada tema para marcas comerciales.

Creación de grupos de prueba

Para realizar los experimentos creamos tres grupos de usuarios, uno para cada una de las estrategias mostradas en la Tabla 4.12. En esta tabla se muestra el número de contactos por usuario, número de comentarios mínimo por contacto y el total de comentarios en la representación del usuario. El *vector de usuario* $\vec{U}$ se cró según la estrategia de cada grupo pero la experimentación realizada en ellos sigue el mismo proceso explicado en la Sección 4.3.1

El **Grupo 1** y y el **Grupo 2** están formados cada uno por 6,000 usuarios elegidos al azar a partir de los usuarios que existen en la muestra que obtuvimos. Estos usuarios cumplen la condición de seguir por lo menos a otros 40 usuarios

Temas Generales			
Tema	Cantidad	Tema	Cantidad
Música	45685	Cine	25849
Deporte	12974	Cultura	9909
Tecnología	9897	Ciencia	4818
Teatro	4418	Religión	3639

Tabla 4.16: Palabras utilizadas para obtener los comentarios relacionados con cada tema general.

Grupo	Usuarios	Contactos	Comentarios	Representación
Grupo1	6000	20	> 400	mínimo 15184
Grupo2	6000	40	> 400	mínimo 32292
Grupo3	400	30	> 0	mínimo 17121

Tabla 4.17: Resumen de los grupos de prueba.

que a su vez han publicado por lo menos 400 comentarios cada uno. Elegimos 400 comentarios por ser una cantidad estadísticamente significativa. Los usuarios elegidos son utilizados en ambos grupos con la diferencia en la utilización de una mayor cantidad de contactos en el Grupo 2. Para el Grupo 1 utilizamos 20 contactos y en el Grupo 2 utilizamos 40 contactos que incluyen a los 20 primeros. De esta forma podemos comparar el efecto que tiene la utilización de un mayor o menor número de contactos al momento de cuantificar el interés.

El **Grupo 3** es un grupo de 400 usuarios elegidos al azar que cumplen la condición de seguir por lo menos a otros 30 usuarios con los cuales haya tenido contacto mediante mención. En el caso de Twitter, la mención se define como la utilización del nombre del usuario en el comentario el cual recibe una notificación para su lectura [106]. Esta muestra, aunque es lo suficientemente grande para ser estadísticamente significativa es pequeña en comparación con los usuarios de los Grupos 1 y 2 debido a que la cantidad de usuarios que interactúan mediante mención mutua en Twitter es mucho menor en nuestra muestra. En Twitter existen usuarios conocidos como *bots*

los cuales podrían ser personas que operan bajo un sueldo o programas de computadora que simulan ser un humano. Creamos este grupo siguiendo la intuición de que si existe una comunicación entre los usuarios mediante mención podemos reducir la probabilidad de que alguno de ellos sea un *bot* ya que la presencia de éstos puede influir en la cuantificación de los intereses del usuario. Existen algoritmos avanzados para la detección automática de *bots* [20], sin embargo la implementación de alguno está fuera del alcance de la presente tesis.

Los grupos de prueba y sus características se resumen en la Tabla 4.17. Para realizar los experimentos, en cada grupo de prueba elegimos al azar un número de pares de usuarios y los separamos en tres categorías según la Tabla 4.18. Para el Grupo 1 y 2 elegimos 12,000 pares de usuarios en cada categoría y para el Grupo 3 elegimos 4,000 pares en cada categoría. En el caso de una relación unidireccional para el Grupo 3 solo nos fue posible obtener 900 pares. Quizá la probabilidad de que los usuarios tengan una relación unidireccional es menor cuando se han mencionado el uno al otro, favoreciendo a la relación bidireccional. La cantidad de pares por grupo y categoría se muestra en la Tabla 4.19.

Tipo	Descripción
Sin relación	Ninguno de los usuarios sigue al otro.
Unidireccional	Un usuario sigue al otro pero éste no lo sigue a él.
Bidireccional	Ambos usuarios se siguen mutuamente.

Tabla 4.18: Tipos de contactos en Twitter.

La probabilidad de que dos usuarios sean contactos según el valor del Interés hacia el mismo tema se calculó para cada grupo de dos formas según la relación unidireccional o bidireccional. La primera consiste en dividir la cantidad de usuarios que tienen una relación unidireccional entre la cantidad de usuarios que no tienen relación (que no son contactos). En la segunda se divide la cantidad de usuarios que tienen una relación bidireccional entre la cantidad de usuarios que no tienen relación. Finalmente calculamos la correlación que existe entre el Interés y la probabilidad de que dos usuarios sean contactos para cada una de las estrategias planteadas para los

Grupo	Categoría	Número de pares
Grupo 1	Sin relación	12000
	Unidireccional	12000
	Bidireccional	12000
Grupo 2	Sin relación	12000
	Unidireccional	12000
	Bidireccional	12000
Grupo 3	Sin relación	4000
	Unidireccional	900
	Bidireccional	4000

Tabla 4.19: Pares de usuarios para la experimentación por cada grupo.

Grupos 1, 2 y 3.

En la siguiente sección analizaremos los resultados.

4.3.2 Resultados

La muestra global se compone de 40,186,542 comentarios los cuales fueron clasificados según su sentimiento utilizando la herrameinta TOM. En la Tabla 4.20 se muestra la cantidad de comentarios para cada clasificación. El 65 % de los comentarios son neutros, 17 % son negativos y 18 % son positivos. En la Sección 4.3.1 mencionamos que utilizaríamos subconjuntos según la clase del sentimiento de los comentarios. Vemos que esto sería conveniente computacionalmente hablando ya que al utilizar solo una clase de sentimiento se podría reducir el tiempo de procesamiento de texto al discriminar los demás comentarios.

Para cada usuario en los tres grupos de prueba obtuvimos el Interés para los 38 temas con los que trabajamos. En la Tabla 4.21 se muestran estadísticas descriptivas del Interés por grupo y subconjunto de comentarios según el sentimiento.

Clasificación	Cantidad	Porcentaje
Positivos	7,221,178	18 %
Negativos	6,753,920	17 %
Neutros	26,211,444	65 %

Tabla 4.20: Sentimiento de los comentarios del repositorio.

Como mencionamos anteriormente, los comentarios de los usuarios elegidos para representar al usuario y evaluar la función $c_u(t)$ pueden variar según el sentimiento; esto es, se puede utilizar una sola categoría de comentarios (positivos, negativos o neutros) para crear el *vector de usuario* $\vec{U}$. De esta forma obtenemos distintos valores del Interés para cada categoría, incluyendo el caso global en donde se utilizan todos los comentarios. Podemos notar que los valores en general son muy pequeños; esto se debe al uso de la función de similitud cosenoidal para calcular la similitud entre el tema y el contenido ($c_u(t)$) la cual requiere que los *vectores de usuario* $\vec{U}$ sean muy similares para obtener valores cercanos a 1.

Grupo	Sentimiento	Mínimo	Máximo	Rango	Promedio	Desviación
Grupo 1	Global	-0.0336	0.0814	0.1149	-0.0001	0.0016
	Negativos	-0.0482	0.0299	0.0634	-0.0003	0.0019
	Neutros	-0.0190	0.0528	0.0864	-0.0001	0.0015
	Positivos	-0.0133	0.2784	0.3120	0.0000	0.0025
Grupo 2	Global	-0.0274	0.0288	0.0562	-0.0001	0.0013
	Negativos	-0.0524	0.0189	0.0463	-0.0002	0.0016
	Neutros	-0.0182	0.0306	0.0580	-0.0001	0.0013
	Positivos	-0.0114	0.1016	0.1290	0.0000	0.0017
Grupo 3	Global	-0.0206	0.0211	0.0417	-0.0003	0.0017
	Negativos	-0.0362	0.0186	0.0392	-0.0004	0.0020
	Neutros	-0.0175	0.0196	0.0402	-0.0003	0.0017
	Positivos	-0.0127	0.0169	0.0375	-0.0002	0.0015

Tabla 4.21: Estadísticas descriptivas del Interés.

Para cada uno de los grupos calculamos las probabilidades de que dos usuarios sean contactos para intervalos acotados que parten de k desviaciones estándar hasta 1 cuando al usuario le agrada el tema ($I > 0$), y de -1 hasta $-k$ desviaciones estandar cuando al usuario le desagrada el tema ($I < 0$). Analizamos ambas partes (agrado y desagrado) de forma separada ya que deseamos observar si una opinión negativa con respecto a un tema de ambos usuarios aumenta la probabilidad de que sean contactos de la misma que una opinión positiva.

Así mismo, en cada grupo hicimos un análisis de la correlación que existe entre la probabilidad de que dos usuarios sean contactos y el Interés por el mismo tema. Para calcular dicha correlación se utilizó un conjunto P de valores discretos del Interés para los que se evaluó la probabilidad de que dos usuarios sean contactos. Para ésto se utilizaron rangos del Interés que varían +/- 0.001 partiendo de cada punto discreto $p_i \in P$. Estos puntos discretos comienzan en 0 y van incrementandose o decrementandose en la cantidad de 0.0005 o -0.0005 (son multiplos de 0.0005). Aquellos puntos para los que existió una cantidad de usuarios mayor o igual a 30 en el rango del Interés $[p_i - 0.001, p_i + 0.001]$, fueron utilizados para el cálculo de la correlación.

En las siguientes subsecciones presentamos los resultados para cada uno de los grupos de prueba.

Resultados y discusión del Grupo 1

La Tabla 4.22 muestra la variación de la probabilidad de que dos usuarios sean contactos conforme su Interés por el mismo tema aumenta. Observamos que a medida que el Interés se hace más positivo o más negativo la probabilidad aumenta. En el caso unidireccional tenemos que cuando el Interés es menor a -3σ (desagrado) la probabilidad alcanza el 74 % utilizando comentarios globales y neutros para calcular $c_u(t)$. Cuando el usuario tiene un agrado hacia el tema ($I > 0$) la probabilidad alcanza el 69 % cuando el Interés es mayor a 3σ y cuando se utilizan comentarios

positivos para calcular $c_u(t)$.

Grupo 1								
	$[-1, \mu - k\sigma]$				$[\mu + k\sigma, 1]$			
Tipo	$k=3$	$k=2$	$k=1$	$k=0$	$k=0$	$k=1$	$k=2$	$k=3$
Unidireccional								
Neutros	**0.74**	0.67	0.61	0.55	0.48	0.52	0.58	0.66
Positivos	0.50	0.73	0.61	0.53	0.48	0.53	0.64	**0.69**
Negativos	0.67	0.65	0.59	0.55	0.49	0.50	0.55	0.64
Global	**0.74**	0.68	0.61	0.55	0.48	0.52	0.58	0.65
Máximo	0.74	0.73	0.61	0.55	0.49	0.53	0.64	0.69
Bidireccional								
Positivos	**0.80**	0.78	0.65	0.57	0.42	0.33	0.41	**0.50**
Neutros	0.77	0.71	0.64	0.59	0.44	0.35	0.35	0.41
Negativos	0.71	0.69	0.63	0.59	0.46	0.35	0.33	0.35
Global	0.75	0.70	0.61	0.59	0.45	0.39	0.43	0.49
Máximo	0.80	0.78	0.65	0.59	0.46	0.35	0.41	0.50

Tabla 4.22: Variación de la probabilidad de ser contactos con respecto al Interés para el Grupo 1.

Llama la atención que cuando ambos usuarios manifiestan un desagrado por el mismo tema su probabilidad de ser contactos tiende a aumentar; mientras que en el caso bidireccional observamos que la probabilidad cuando los usuarios muestran un agrado alto hacia el tema no supera el 50 % en ningún caso. Esto quiere decir que en las relaciones de usuarios más estrechas (ya que ambos siguen al otro) podría tener un mayor peso el desagrado por los mismos temas lo que hace que sean contactos. Se podría investigar en un trabajo futuro si en los grupos más cohesivos el desagrado por el mismo tema los une más o si este resultado es un efecto provocado por la interface de Twitter (nuestro caso de estudio). Las Figuras 4.6 y 4.7 muestran gráficas de la probabilidad de ser contactos conforme varía el Interés para el Grupo 1 en los casos unidireccional y bidireccional.

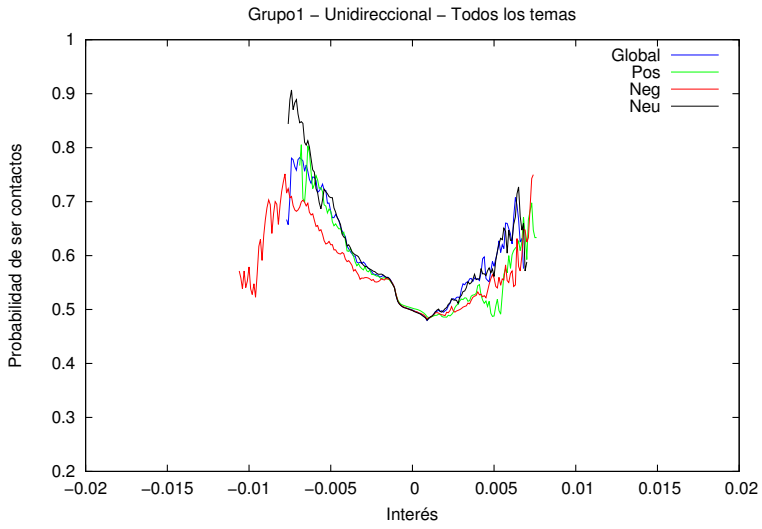


Figura 4.6: Probabilidad de ser contactos según el Interés para el Grupo 1 y relación unidireccional.

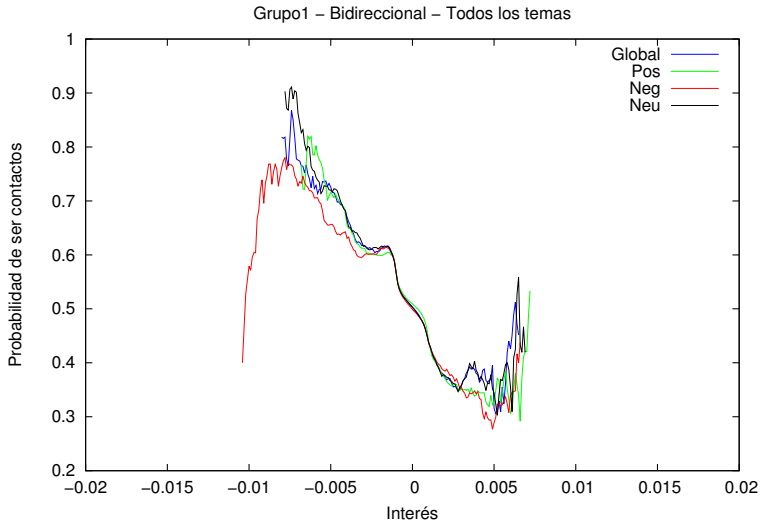


Figura 4.7: Probabilidad de ser contactos según el Interés para el Grupo 1 y relación unidireccional.

Grupo 1				
Sentimiento	Unidireccional		Bidireccional	
	$I < 0$	$I > 0$	$I < 0$	$I > 0$
Positivos	-0.9753	0.8392	-0.9703	-0.5872
Neutros	-0.9710	0.8643	-0.9644	-0.0566
Negativos	-0.5587	0.9336	-0.6986	-0.6908
Global	-0.9575	0.8662	-0.9825	-0.2431
Promedio	-0.8656	0.9235	-0.9040	-0.3944

Tabla 4.23: Correlación de la probabilidad de ser contactos y el Interés según el tipo de sentimiento para el Grupo 1.

Los resultados de un análisis de correlación realizado para el agrado y desagrado en los casos unidireccional y bidireccional se muestra en la Tabla 4.23. En general observamos correlaciones indicativas de una relación fuerte excepto cuando los usuarios muestran un agrado por el tema ($I > 0$) en el caso bidireccional. En este caso la correlación es negativa. Esto sugiere que conforme el agrado por un tema aumenta es menos probable que los usuarios sean contactos mutuos (relación bidireccional). En trabajos futuros podría investigarse si esto se debe a cuestiones psicológicas de las personas.

En general observamos que la forma en que se eligen los comentarios que servirán para crear la representación del usuario tiene un impacto en la cuantificación del interés. Utilizar comentarios positivos obtuvo los mejores resultados en el caso unidireccional y bidireccional cuando ambos usuarios muestran un agrado ($I > 0$). Cuando ambos usuarios muestran un desagrado por el tema, la utilización de comentarios globales y neutros obtuvieron los mejores resultados en el caso unidireccional y los positivos para el caso bidireccional. Sin embargo, en este caso los neutros y negativos también tienen probabilidades altas (77 % y 71 % respectivaemnte).

Resultados y discusión del Grupo 2

La Tabla 4.24 muestra la variación de la probabilidad de que dos usuarios sean contactos conforme su Interés por el mismo tema aumenta. Observamos que a medida que el Interés se hace más positivo o más negativo la probabilidad aumenta al igual que para el Grupo 1. En el caso unidireccional tenemos que cuando el Interés es menor a -3σ la probabilidad alcanza el 71 % para comentarios globales y neutros. Cuando los usuarios manifiestan un agrado hacia los temas la probabilidad alcanza también un 71 % cuando el Interés es mayor a 3σ.

Grupo 2								
	$[-1, \mu - k\sigma]$				$[\mu + k\sigma, 1]$			
Tipo	$k = 3$	$k = 2$	$k = 1$	$k = 0$	$k = 0$	$k = 1$	$k = 2$	$k = 3$
Unidireccional								
Global	**0.71**	0.64	0.59	0.54	0.49	0.56	0.66	**0.71**
Negativos	0.68	0.64	0.58	0.55	0.49	0.55	0.65	0.64
Neutros	**0.71**	0.64	0.58	0.54	0.49	0.56	0.65	0.70
Positivos	0.62	0.62	0.58	0.53	0.49	0.56	0.64	0.70
Máximo	0.71	0.64	0.59	0.55	0.49	0.56	0.66	0.71
Bidireccional								
Global	0.75	0.70	0.61	0.59	0.45	0.39	0.43	0.49
Negativos	0.71	0.70	0.61	0.59	0.46	0.40	0.44	0.50
Neutros	**0.77**	0.70	0.61	0.59	0.45	0.39	0.42	0.45
Positivos	0.73	0.75	0.61	0.57	0.44	0.37	0.44	**0.56**
Máximo	0.77	0.75	0.61	0.59	0.46	0.40	0.44	0.56

Tabla 4.24: Variación de la probabilidad de ser contactos con respecto al Interés para el Grupo 2.

En el caso bidireccional sucede algo parecido al Grupo 1, la probabilidad llega a un 56 % cuando los usuarios manifiestan un agrado hacia el tema ($I > 0$); sin embargo, cuando los usuarios muestran un desagrado hacia el tema la probabilidad

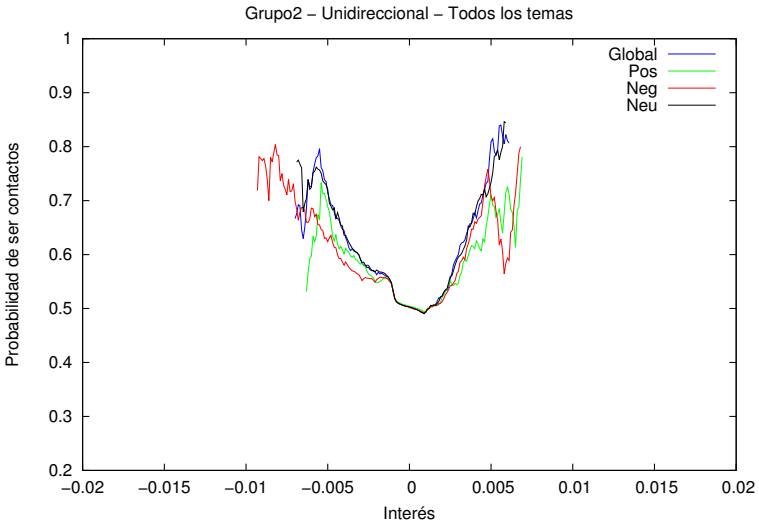


Figura 4.8: Probabilidad de ser contactos según el Interés para el Grupo 2 y relación unidireccional.

alcanza un 77 %. Aquí el patron del Grupo 1 se repite, por lo que nuevamente surge la pregunta del efecto que tiene en los usuarios el compartir un desagrado con respecto a un tema y su posible origen debido a la interface de Twitter o psicología de las personas. Las Figuras 4.8 y 4.9 muestran gráficas de la probabilidad de ser contactos conforme varía el Interés para el Grupo 2 en los casos unidireccional y bidireccional.

Grupo 2				
Sentimiento	Unidireccional		Bidireccional	
	$I < 0$	$I > 0$	$I < 0$	$I > 0$
Positivos	-0.8641	0.8933	-0.6691	0.2137
Neutros	-0.9368	0.9829	-0.9473	-0.3104
Negativos	-0.9753	0.8459	-0.9427	-0.1284
Global	-0.8177	0.9720	-0.8308	0.2855
Promedio	-0.8985	0.9235	-0.8475	0.0151

Tabla 4.25: Correlación de la probabilidad de ser contactos y el Interés según el tipo de sentimiento para el Grupo 2.

Los resultados de un análisis de correlación realizado para el agrado y desagra-

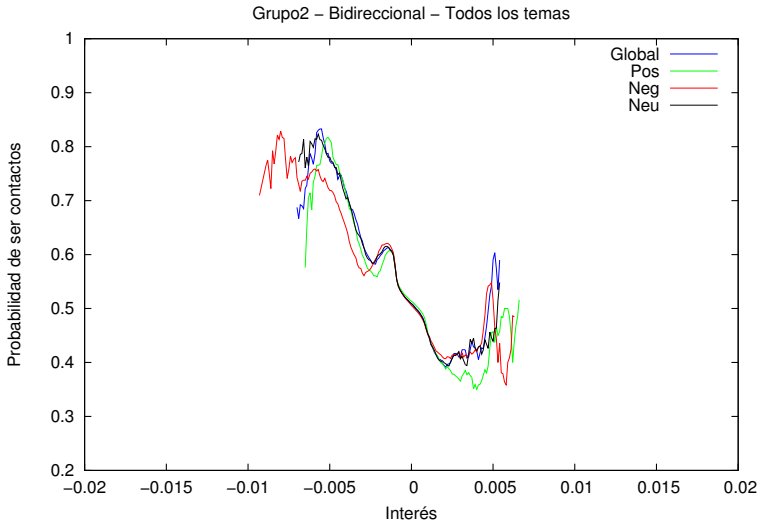


Figura 4.9: Probabilidad de ser contactos según el Interés para el Grupo 2 y relación unidireccional.

do en los casos unidireccional y bidireccional se muestra en la Tabla 4.25. Al igual que en el Grupo 1, observamos correlaciones indicativas de una relación fuerte excepto cuando los usuarios muestran un agrado por el tema ($I > 0$) en el caso bidireccional. En este caso la correlación es negativa cuando se utilizan comentarios neutros y negativos para crear la representación del usuario. Sin embargo, la correlación es positiva (aunque débil) cuando se utilizan comentarios positivos. Aquí surge la duda si la utilizaciń de un mayor numero de conactos del usuario podría seguir cambiando los resultados en cuanto a la probabilidad calculada mediante las relaciones bidireccionales. Esto podría investigarse en un trabajo futuro.

Al igual que en el Grupo 1 observamos que la forma en que se eligen los comentarios que servirán para crear la representación del usuario tiene un impacto en la cuantificación del interés. Utilizar comentarios positivos obtuvo los buenos resultados en el caso unidireccional y bidireccional cuando ambos usuarios muestran un agrado ($I > 0$). Cuando ambos usuarios muestran un desagrado por el tema, la utilización de comentarios globales y neutros obtuvieron los mejores resultados en el caso unidireccional y los neutros para el caso bidireccional. En general se obtuvieron probabilidades altas, mayores al 70 %.

Resultados y discusión del Grupo 3

La Tabla 4.26 muestra la variación de la probabilidad de que dos usuarios sean contactos conforme su Interés por el mismo tema aumenta. En el caso unidireccional notamos una diferencia significativa con respecto a los resultados del Grupo 1 y 2. Cuando los usuarios manifiestan un agrado ($I > 0$) por el mismo tema la probabilidad alcanza un 71 %, sin embargo, cuando los usuarios manifiestan un desagrado ($I < 0$) por el tema la probabilidad alcanza un 52 %,

Grupo 3								
	$[-1, \mu - k\sigma]$				$[\mu + k\sigma, 1]$			
Sentimiento	$k = 3$	$k = 2$	$k = 1$	$k = 0$	$k = 0$	$k = 1$	$k = 2$	$k = 3$
Unidireccional								
Global	**0.52**	0.45	0.50	0.48	0.53	0.56	0.61	0.61
Negativos	**0.52**	0.46	0.49	0.48	0.52	0.56	0.62	**0.71**
Neutros	0.51	0.45	0.49	0.48	0.53	0.56	0.62	0.70
Positivos	0.51	0.47	0.50	0.48	0.53	0.58	0.61	0.65
Máximo	0.52	0.47	0.50	0.48	0.53	0.58	0.62	0.71
Bidireccional								
Global	**0.62**	0.60	0.57	0.55	0.49	0.48	0.49	0.53
Negativos	**0.62**	0.59	0.56	0.55	0.49	0.48	0.52	**0.60**
Neutros	**0.62**	0.59	0.56	0.55	0.49	0.48	0.49	0.51
Positivos	0.60	0.59	0.57	0.55	0.49	0.48	0.50	0.53
Máximo	0.62	0.60	0.57	0.55	0.49	0.48	0.52	0.60

Tabla 4.26: Variación de la probabilidad de ser contactos con respecto al Interés para el Grupo 3.

En el caso bidireccional las probabilidades cuando los usuarios muestran un agrado ($I > 0$) o desagrado ($I < 0$) alcanzan valores de 60 % y 62 % respectivamente. Vemos que aunque se observa una tendencia creciente, las probabilidades son generalmente más bajas que en los Grupos 1 y 2. La máxima probabilidad de que

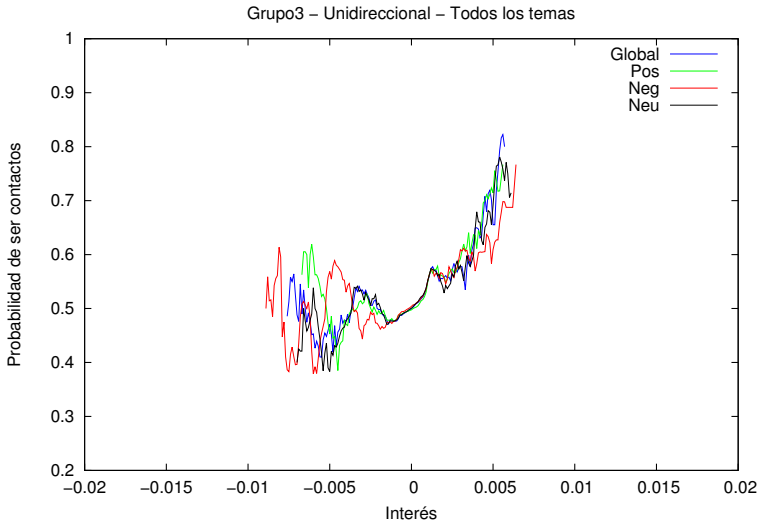


Figura 4.10: Probabilidad de ser contactos según el Interés para el Grupo 3 y relación unidireccional.

los usuarios sean contatos en el Grupo 3 se da en el caso unidireccional cuando los usuarios manifiestan un agrado hacia el mismo tema. Comparando estos resultados con los de los Grupos 1 y 2 podemos notar que la elección de los usuarios de forma sistemática (no aleatoria) tiene un impacto significativo al cuantificar el Interés. Las Figuras 4.10 y 4.11 muestran gráficas de la probabilidad de ser contactos conforme varía el Interés para el Grupo 3 en los casos unidireccional y bidireccional.

Grupo 3				
Sentimiento	Unidireccional		Bidireccional	
	$I < 0$	$I > 0$	$I < 0$	$I > 0$
Positivos	-0.3976	0.9389	-0.7953	0.4417
Neutros	0.3994	0.9150	-0.8442	0.4466
Negativos	0.0650	0.8873	-0.7929	0.4766
Global	0.1638	0.8875	-0.5405	0.5881
Promedio	0.0576	0.9072	-0.7432	0.4883

Tabla 4.27: Correlación de la probabilidad de ser contactos y el Interés según el tipo de sentimiento para el Grupo 3.

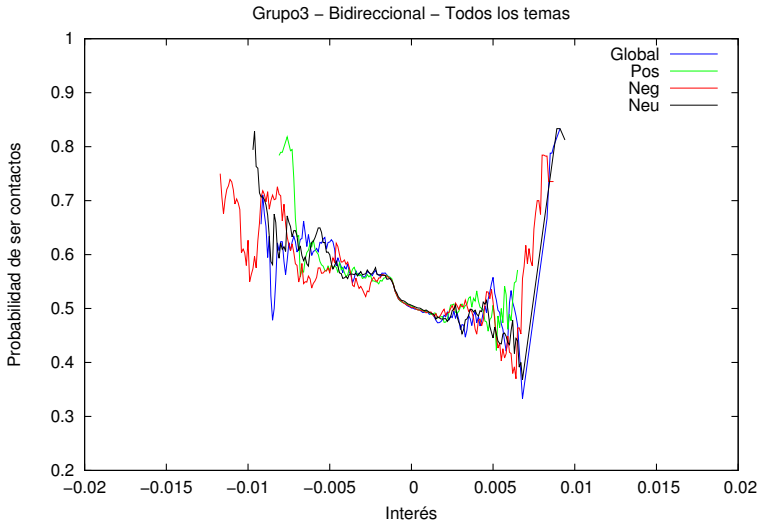


Figura 4.11: Probabilidad de ser contactos según el Interés para el Grupo 3 y relación bidireccional.

Los resultados de un análisis de correlación realizado para el agrado y desagrado en los casos unidireccional y bidireccional se muestra en la Tabla 4.27. Podemos observar que las correlaciones cuando los usuarios muestran un agrado hacia el tema ($I > 0$) indican una relación fuerte y moderada para el caso unidireccional y bidireccional respectivamente. Esto contrasta con los casos del Grupo 1 y 2, ya que la correlación en estos es débil o incluso negativa en el caso bidireccional.

Al igual que en los Grupos 1 y 2 observamos que la forma en que se eligen los comentarios que servirán para crear la representación del usuario tiene un impacto en la cuantificación del interés. Utilizar comentarios positivos obtuvo los mejores resultados en el caso unidireccional. Los comentarios neutros tienen una correlación mayor en el caso bidireccional cuando los usuarios muestran un desagrado por el mismo tema ($I < 0$) mientras que con el uso de todos los comentarios se obtienen mejores resultados cuando los usuarios muestran un agrado por el mismo tema ($I > 0$). No obstante, la utilización de comentarios positivos, negativos o neutros produce resultados similares en el caso bidireccional por lo que podría ser útil utilizar un conjunto de comentarios positivos para representar al usuario siguiendo la estrategia de selección de contactos utilizada en el Grupo 3.

4.3.3 Discusión general de los resultados

En los Grupo 1 y 2 vemos que el Interés tiene una correlación fuerte en el caso unidireccional tanto para el agrado ($I > 0$) como el desagrado ($I < 0$); mientras que el caso bidireccional la correlación más fuerte se da en el caso del desagrado ($I < 0$) solamente. En el Grupo 3 las correlaciones son más fuertes en el caso bidireccional y cuando los usuarios tienen un agrado ($I > 0$) por el mismo tema en el caso unidireccional. Vemos que el hecho de que dos usuarios compartan un agrado o desagrado por un tema aumenta la probabilidad de que los usuarios sean contactos.

En el caso unidireccional, para el Grupo 1 tenemos que cuando el Interés es menor a -3σ la probabilidad alcanza el 74 % para comentarios globales y neutros. En el lado positivo del Interés la probabilidad alcanza el 69 % cuando el Interés es mayor a 3σ. Para el Grupo 2 tenemos que si el Interés es menor a -3σ o mayor a 3σ la probabilidad alcanza el 71 % para comentarios globales y neutros. En el caso del Grupo 3, cuando el Interés es menor a 3σ la probabilidad llega a 52 % mientras que cuando el Interés es mayor a 3σ la probabilidad alcanza el 71 %.

En estos grupos, la utilización de comentarios neutros y globales alcanza las probabilidades más altas cuando se trabaja con el desagrado de los usuarios. Sin embargo, al trabajar con el agrado, los comentarios positivos alcanzan probabilidades de 70 % para el Grupo 2, 69 % para el Grupo 1 y 71 % para el Grupo 3.

En el caso del Grupo 2 podría ser preferible utilizar los comentarios positivos al trabajar con el agrado ya que la cantidad de comentarios de éstos es mucho menor. El Grupo 3 es un caso especial por la forma en que se seleccionaron los comentarios de los usuarios para crear las representaciones. En éstos se incluyen sólo comentarios de contactos con los que haya interactuado mediante mención. Observamos que cuando el Interés es menor a -3σ la probabilidad se mantiene cercana al 50 % mientras que cuando es mayor a 3σ alcanza el 71 %. Esto quiere decir que un agrado de los mismos temas ($I > 0$) aumenta la probabilidad de que los usuarios que tienen interacciones

entre ellos sean contactos. Mientras que en los Grupos 1 y 2, las relaciones se forman independientemente cuando hay un Interés alto ya sea negativo o positivo por los mismos temas.

Para el caso de la relación bidireccional para el Grupo 1 (ver Tabla 4.22) tenemos que cuando el Interés es menor a -3σ la probabilidad alcanza el 80 % mientras que cuando es mayor a 3σ llega a un 50 %, ambos utilizando los comentarios positivos. Para el Grupo 2, cuando el Interés es menor a -3σ la probabilidad alcanza un 77 % con los comentarios neutros mientras que cuando es mayor a 3σ llega a un 56 % con los comentarios positivos. En el Grupo 3, cuando el Interés es menor a -3σ la probabilidad alcanza un 62 % y cuando es mayor a 3σ llega a un 60 %, ambas utilizando los comentarios negativos. En los Grupos 1 y 2 observamos que el Interés sirve mejor para predecir si dos usuarios son contactos cuando tienen un desagrado por los mismos temas, mientras que, aunque tengan un agrado, no implica una probabilidad alta de que sean contactos de forma bidireccional. El caso bidireccional es interesante ya que ambos usuarios han manifestado un interés en el otro. En el Grupo 3 la probabilidad aumenta cuando los usuarios comparten un agrado, llegando a un 60 % cuyo valor es preferible al de 56 % o 50 % de los Grupos 1 y 2.

Para el caso unidireccional observamos que la correlación para los tres grupos es alta en el caso de agrado por el mismo tema; con valores arriba de 0.86, pero en caso de desagrado, la correlacion es más baja para el Grupo 3, con un promedio de 0.0576 en comparación con -0.8656 y -0.8985 para los Grupos 1 y 2 respectivamente. Con base en las correlaciones mostradas podemos decir que la utilización de las estrategias para definir al usuario utilizadas por los Grupos 1 y 2 podrían ser preferibles para representar el Interés. La diferencia en este caso entre los Grupos 1 y 2 es pequeña, pero dado que para el Grupo 1 se procesa la mitad del texto que para el Grupo 2 podría ser preferible sobre la estrategia del Grupo 1.

En el caso bidireccional observamos que la correlación para el Grupo 3 es preferible a la de los los Grupos 1 y 2 en la parte del agrado por el mismo tema, para el cual los Grupos 1 y 2 tienen correlaciones bajas (0.0151 en el Grupo 2);

incluso una de ellas es negativa (-0.3944 en el Grupo 1). La correlación negativa del Grupo 1 indica que a medida que aumenta el Interés diminuye la probabilidad de que los usuarios sean contactos por lo que esta estrategia no sería recomendable para predecir si dos usuarios son contactos o no. Sin embargo, para los Grupos 1 y 2 existe una correlación alta en la parte del desagrado ($I > 0$) con promedios de -0.9040 y -0.8475 respectivamente mayor a la del Grupo 3 de -0.7432.

Con base en los datos mostrados, para crear un sistema de detección de Interés alto con relación a una marca o producto por parte de los usuarios, se podría utilizar la estrategia planteada en el Grupo 3 siempre que sea posible mediante la utilización de comentarios positivos provenientes de contactos con los que ha interactuado mendiante mención y utilizando el agrado ($I > 0$) para seleccionar a los usuarios que sobrepasen un cierto limite θ. Si el usuario no tiene contactos cuya relación es bidireccional y haya tenido alguna interacción se podría utilizar alguna estrategia como la del Grupo 1 que utiliza solo 20 contactos al azar pudiendose utilizar también solo los comentarios positivos.

En el caso de tener la intención de crear un sistema de recomendación de usuarios una posible forma de hacerlo sería a través del desagrado ($I < 0$) por los mismos temas utilizando la estrategia planteada por el Grupo 1 la cual presenta una correlación alta en ambos tipos de relaciones (unidireccional y bidireccional) con -0.9316 y -0.9567 respectivamente utilizando solo los comentarios positivos.

En los tres grupos se utilizaron comentarios de los contactos del usuario para crear su representación y, posteriormente, estos comentarios fueron evaluados utilizando la herramienta TOM. Por simplicidad, en la cuantificación del interés se llevó a cabo la evaluación de todos los comentarios sin excepción. Dados los resultados favorables con respecto al Interés, podemos asumir que el hecho de no eliminar comentarios en otros idiomas para crear la representación del usuario no afecta demasiado al cuantificar el interés en un tema,— dichos comentarios serían clasificados como neutros por TOM ya que no identificaría palabras con sentimiento. En el caso de que un usuario utilice dos o más idiomas con frecuencia, quizá sea conveniente

aplicar un método de detección de idioma previo a la cuantificación del interés o modificar a TOM para incluir la evaluación de dichos comentarios. Esto será analizado en un trabajo futuro.

Durante la evaluación de TOM se encontró que en el 50.17 % de los comentarios del conjunto de prueba C_2 los evaluadores no coincidieron en la clasificación por lo que no se encuentran dentro de ninguna categoría. Estos comentarios podrían manifestarse en forma de ruido en la cuantificación del interés ya que podrían generar falsos positivos, negativos o neutros. En un trabajo futuro, para medir esta cantidad de ruido podríamos plantear la realización de una comparación entre la evaluación de TOM en estos comentarios y el acuerdo o desacuerdo en dicha evaluación por parte de un cierto grupo de evaluadores. Así podríamos observar si la respuesta de TOM es favorecida en caso de indecisión en cuanto al sentimiento.

4.4 Recomendación de temas

Utilizamos el criterio planteado en la sección anterior para recomendación de temas utilizando comentarios positivos para la representación del usuario y la estrategia de selección de contactos del Grupo 3 con valores del Interés mayores a 2σ (0.0027) para hacer la recomendación. La cantidad de usuarios a los que se les podría recomendar el tema se muestran en la Figura 4.12.

En la Tabla 4.28 se muestran 18 comentarios de un usuario cuyo Interés por la marca “TV Azteca” es alto. Algunas menciones fueron eliminadas para proteger su identidad, se dejaron aquellas referencias de personas públicas. Así mismo, la Tabla 4.29 muestra las 60 palabras que obtuvieron el valor TF-IDF más alto para la misma marca.

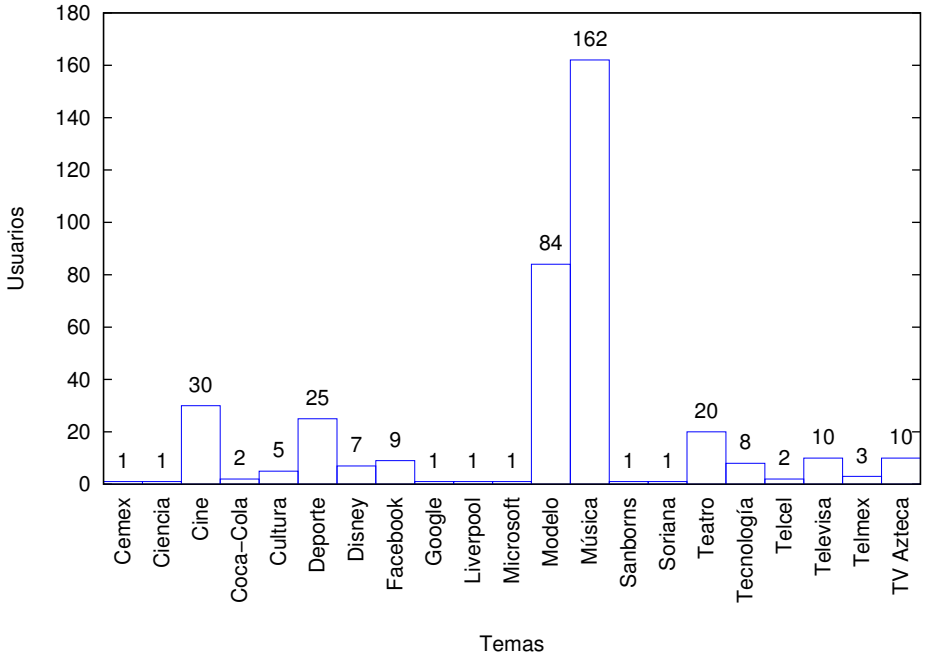


Figura 4.12: Usuarios con Interés en el tema.

Espero que no!! Y que le guste a todo el mundo!
Saludos Analeli!! Bendiciones
Saliendo del Circo de la Cafetería con mi compa @regalito_tv con
grandes planes y sorpresas para ustedes. Estén pendientes!!
Esa rolita sí que me gusta! Lo malo es que la hice yo.
Eh! Yo quiero ir! Ya no me invitan! Qué gachos... Jejejeje.
Saludos a la familia compa!
Buenis días a todos! Madrugué por culpa de una maldita pesadilla
y mejor me vine a correr a la arboleda. Y terminé, así que un bañito!
Buenos días @lizbriones @OMARCHAPARRO. Pura gente disciplinada.
Buenos días @teregarcia1931 @ladivademexico .
Buenos días @soyomareldeclub @MikyCopeton
pero soy compa de todos! Soy González!
ya está! Me acuerdo cuando andaba bien necio con El planeta de los
simios decía “vamos a ver la de changos changos”. Jejeje.
@EDSONZUNIGA Qué bueno que regresó compa! A ver si le marco
para saludarlo. Un fuerte abrazo. Elías
@OMARCHAPARRO Compadrito! Revise su email porque ya le mandé su
encargo. Un gran abrazo y espero que le guste.
Esta es una convocatoria para que me digan a dónde llevar a pasear a
mi novia hoy en MTY. Descartados el cine, boliche, cenar, Santa Lucía!!
Esa es muy buena! Lo malo es que ella vive en Sta. Catarina y nos queda
lejos. Pero muchas gracias hermano!
Ya fuimos, nos subimos y hasta cenamos en El Lingote la mejor comida que
jamás haya probado. Saludos compa! Mas sugerencias
Neta que soy bien fan de @ladivademexico !
@fernandolozano ojalá que ganen hoy y mañana.
Saludos compadrito. Un gran abrazo!

Tabla 4.28: Comentarios de ejemplo de un usuario con Interés alto en TV Azteca.

azteca	televisora	narran
tv.azteca	duopolio	deportv
monopoliza	diferido	maywea
diferidos	mexicanosen	myriam
noootable	espn	gost
noreste	multará	comentarista
empres	cronistas	redzone
4121	bezares	penalizado
enchila'o	andre	sigal
narradores	excepsion	marin
encuetran	balonazos	necaxa
telerisa	informarles	desenmascaró
multimedios	convencen	smackdown
comentaristas	televisa	azteca.wmv
martinolli	futas	moreliavs
diferida	narradadores	eikona
extranormal	andré	caibien
laborar	martinoli	mynetwork
narracion	ventaneando	azteca.mp4
razanchulamque	partidod	minipula

Tabla 4.29: Lista de 60 palabras que obtuvieron el valor TF-IDF más alto para la marca TV Azteca.

Capítulo 5

Conclusiones

En esta tesis abordamos el problema de calcular el Interés de un usuario en un tema mediante minería de texto y análisis de sentimiento. Definimos formalmente el Interés de un usuario en un tema como una función f del contenido del usuario $c_u(t)$ y del sentimiento expresado en el mismo $s_u(t)$, cuyo valor es un indicador del agrado, desagrado o indiferencia con respecto al tema. Utilizamos minería de texto para obtener $c_u(t)$, el cual mide la similitud del usuario con el tema, y análisis de sentimiento en el contenido para obtener $s_u(t)$, la cual es el sentimiento del usuario con respecto al tema. En los experimentos de esta tesis utilizamos la multiplicación como función de cuantificación f ($f(c_u(t), s_u(t)) = c_u(t) \cdot s_u(t)$), sin embargo, otras funciones podrían ser investigadas.

Las funciones $c_u(t)$ y $s_u(t)$ deben ser elegidas de acuerdo al entorno en el que se esté trabajando con el fin de representar en forma adecuada el interés del usuario. Este concepto puede ser aplicado tanto en redes sociales como en otros ámbitos donde exista una interacción del usuario a través de comentarios como en el caso de los microblogs. Tanto el contenido como el sentimiento pueden variar en la forma dependiendo del entorno y éstos pueden ser utilizados para cuantificar el interés.

Para la cuantificación del interés propusimos utilizar una representación del usuario creada a partir de comentarios de los contactos del usuario en un microblog. Analizamos tres posibilidades de representación del usuario, las cuales se listan en la Tabla 4.12. Además analizamos la posiblidad de utilizar comentarios de una sola clase de sentimiento para crear la representación como se explicó en la Sección 4.3.1. De

igual forma, propusimos utilizar comentarios dentro del microblog para representar al tema.

Para lograr el objetivo de hacer análisis de sentimiento en español sobre el contenido del usuario creamos una herramienta llamada TOM, la cual probó tener una precisión considerable y competitiva con respecto a otras herramientas existentes en el estado del arte. Utilizamos dicha herramienta para calcular el sentimiento del usuario con respecto al tema $s_u(t)$. También utilizamos a TOM para elegir los comentarios que se utilizarían para representar al usuario con base en un sentimiento a manera de filtro. TOM podría ser utilizado en otras aplicaciones que tengan que ver con minería de texto y análisis de sentimiento. Así mismo, TOM puede ser mejorado y estudiado por otros investigadores.

En general obtuvimos que la probabilidad de que dos usuarios sean contactos en el microblog aumenta cuando tienen un Interés muy positivo o muy negativo en el tema; lo cual es congruente con el trabajo de Feld sobre relaciones en las organizaciones [33] y el principio de homofilia en redes sociales [69]. La correlación obtenida de dicha probabilidad y el Interés es mayor a 0.80 en la mayoría de los casos. También observamos que utilizar comentarios de una clase de sentimiento para representar al usuario podría ser suficiente para calcular el Interés. Esto podría ser útil durante el cálculo (computacionalmente hablando) ya que es una cantidad de información menor y por lo tanto el tiempo de procesamiento será menor también.

Podemos concluir que sí fue posible utilizar análisis de sentimiento en español con el propósito de cuantificar el interés ya que fuimos capaces de crear y utilizar la herramienta TOM con dicho propósito. Utilizamos los vectores de valores TF-IDF para representar a los usuarios. Por tanto, también logramos aplicar minería de texto en español con el mismo fin. Vimos que los usuarios sí se pueden representar mediante los comentarios de sus contactos y que esto ayuda a cuantificar el interés. Mediante TOM se pudieron crear diferentes representaciones de los usuarios a través de la elección de comentarios que pertenecen a una misma categoría de sentimiento. Utilizamos comentarios del microblog para crear representaciones de los

temas automáticamente. Finalmente, nos fue posible cuantificar el interés utilizando las funciones $c_u(t)$ y $s_u(t)$ lo cual comprueba la hipótesis principal de la tesis. Las siguientes son las principales contribuciones de este trabajo:

- Modelo de cuantifiación del interés con base en el contenido y el sentimiento.
- La herramienta TOM, que clasifica comentarios en español como positivos, negativos o neutros; y su diccionario, el cual incluye el léxico, frases y modificadores de valencia.
- Utilización de comentarios de los contactos del usuario para crear su representación (ver Tabla 4.12).
- Creación de representaciones de un tema apartir de comentarios públicos de un microblog.
- Utilización del sentimiento para elegir los comentarios de los contatos.

5.0.1 Trabajo futuro

Como ya hemos mencionado, la probabilidad de que dos usuarios sean contactos en el microblog aumenta cuando ambos tienen un Interés muy positivo o muy negativo en el mismo tema. La utilización de comentarios de una cierta clase de sentimiento para crear la representación del usuario puede ayudar a que el procesamiento sea más rápido; sin embargo, en un trabajo futuro podría investigarse el impacto que tiene esto para representar los intereses del usuario o en la obtención de un conjunto de temas del agrado o desagrado del usuario.

Para los Grupos 1, 2 y 3, la correlación de la probabilidad de que sean contactos con el interés por el mismo tema es alta cuando se trata de un agrado por los mismos temas para el caso unidireccional. Sin embargo, para el caso bidireccional, la correlación es alta cuando se trata de un desagrado por el mismo tema. Se podría investigar utilizando otros microblogs o redes sociales si en los grupos más cohesivos,

el desagrado por los mismos temas tiene mayor importancia al momento de formar lazos sociales. Así mismo, en el caso unidireccional (en donde un usuario recibe información del otro), podríamos investigar si la relación se debe a que los usuarios desean recibir información o comentarios positivos acerca de sus temas de interés.

La cuantificación del interés realizada en esta tesis obtiene un valor que representa el nivel de agrado, desagrado e indiferencia con respecto a un tema. Sin embargo, para crear un sistema que obtenga automáticamente los usuarios que con mayor probabilidad aceptarían probar un nuevo producto o incluso comprarlo, es necesario determinar un nivel θ, que sirva para clasificar a los usuarios que están dentro de la categoría de posibles compradores. Éste es un trabajo futuro que podría realizarse utilizando productos y usuarios en un escenario que permita validar si probaron o compraron el producto. En otro experimento se podrían utilizar los temas con los que trabajamos en esta tesis para mostrarle a un grupo de usuarios el Interés que resultó de aplicar nuestro método y validar los resultados mediante calificaciones asignadas por ellos.

La experimentación realizada podría llevarse a cabo en otros ambientes donde los usuarios interactúan, como es el caso de redes sociales como Facebook, o Tuenti; incluso se podrían repetir los experimentos en *blogs* de usuarios.

En esta tesis se trabajó con los contactos del usuario para crear una representación que permitiera comparar al usuario con el tema y evaluar su sentimiento. Como mencionamos anteriormente, se utilizaron tres estrategias para elegir dichos usuarios las cuales se encuentran en la Tabla 4.12. Otras estrategias se podrían utilizar para elegirlos; por ejemplo, una vez que se tenga información de usuarios que ya han mostrado un interés en un tema (producto o servicio) con anterioridad, podría ser utilizada para cuantificar el interés de sus contactos hacia dicho tema. En el caso de un producto o servicio, podría ser útil considerar información de los contactos del usuario que ya han tenido un acercamiento con dicho producto o si ya lo han comprado.

El método propuesto para la cuantificación del interés no depende del idioma con el que se esté trabajando. Se podría ampliar la herramienta TOM para procesar también el sentimiento en otros idiomas o utilizar alguna otra para el idioma específico determinado previamente a partir del comentario; ésto sería útil en ciudades donde se utilice más de un idioma con mucha frecuencia.

Por ahora, la precisión de TOM al evaluar el sentimiento es aceptable; sin embargo, todavía se puede mejorar. Para obtener datos confiables del Interés se necesita una cantidad suficientemente grande de comentarios que permita tener la certeza de que el sentimiento con respecto al tema es positivo, negativo o neutro. Conforme se avance en la obtención de una mejor precisión al clasificar los comentarios de TOM, podrían repetirse los experimentos.

El léxico de TOM aún puede mejorarse también. Se podrían utilizar métodos de aprendizaje supervisado que sean capaces de detectar palabras que se utilizan para denotar sentimientos (como por ejemplo, los adjetivos calificativos); utilizando como conjunto de entrenamento y de prueba parte del léxico con el que ya contamos. La determinación de pesos (o carga emotiva) de las palabras también fue realizada manualmente. En este caso se podría utilizar de nueva cuenta algún metodo de aprendizaje supervizado que determine dichos pesos, tales como: máquinas de vectores de soporte (SVM) o redes neuronales. El léxico de SentiWordNet 3.0 [4] en inglés podría ser traducido al español y adaptado al léxico de TOM. Dado que SentiWordNet asigna tres cargas (positiva, negativa y neutra) a cada palabra, un método para determinar el peso final tendrá que ser utilizado antes de agregar dichas palabras al léxico de TOM.

Como mencionamos en la Sección 4.2.2, el corrector de ortografía que utiliza TOM es muy simple. Otros métodos como el que propusimos en dicha sección, el cual utilizaría un cambio sistemático de letras que tengan alguna similitud hasta encontrar la palabra en el léxico y obtener su peso, podría ser utilizado. Dicha similitud podría ser a partir del sonido de las letras (como en el ejemplo de la palabra “kancion” y “canción”).

La lista de modificadores de valencia que utiliza TOM puede ser ampliada. Cierta información proveniente de un analizador de discurso podría ser útil para determinar el alcance del modificador. En los experimentos utilizamos rangos de palabras que son afectadas por los modificadores. Estos rangos incluían una palabra hacia atras o hacia adelante dependiendo del modificador. Este tipo de reglas introdujo algunos errores en la clasificación como se vio en la Tabla 4.10. La información del analizador de discurso podría ser útil para determinar si se trata o no de un modificador de valencia y el alcance del mismo.

Para mejorar la precisión de TOM se podría incluir un proceso de desambiguación de las palabras. Vimos que en el caso de la palabra "Vale" en un comentario de la Tabla 4.10, fue clasificado incorrectamente como negativo ya que no se trataba de una expresión como "Me vale". Para la desambigüación se podría utilizar también un analizador de discurso. Freeling es un analizador de texto en español que se podría utilizar con este fin [78].

Como vimos en la Sección 4.2.1, la eliminación de los sutantivos mediante el método propuesto no fue útil en la clasificación del sentimiento. Se podría seguir investigando este tratamiento mediante un análisis más profundo de las reglas del lenguaje y un analizador de discurso como Freeling. Ciertas frases como "el amor(+1) es feo(-1)", por ahora son clasificadas como neutras por TOM. Eso hace evidente la necesidad de detectar el sustantivo o sujeto que es de lo que se habla.

Bibliografía

[1] Adomavicius, G. y A. Tuzhilin, «Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions», *Knowledge and Data Engineering, IEEE Transactions on*, **17**(6), págs. 734–749, 2005.

[2] Aggarwal, C. y C. Zhai, «An Introduction to Text Mining», *Mining Text Data*, págs. 1–10, 2012.

[3] Alec, B. L. H., Go; Richa, «Twitter Sentiment Classification using Distant Supervision», , 2010.

[4] Baccianella, S., A. Esuli y F. Sebastiani, «Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining», en *Seventh conference on International Language Resources and Evaluation, Malta. Retrieved May*, tomo 25, pág. 2010, 2010.

[5] Baeza-Yates, R., B. Ribeiro-Neto *et al.*, *Modern information retrieval*, tomo 463, ACM press New York., 1999.

[6] Balabanović, M. y Y. Shoham, «Fab: content-based, collaborative recommendation», *Communications of the ACM*, **40**(3), págs. 66–72, 1997.

[7] Balahur, A. y A. Montoyo, «Applying a culture dependent emotion triggers database for text valence and emotion classification», *Procesamiento del lenguaje natural*, **40**, págs. 107–114, 2008.

[8] BALDUZZI, M., C. PLATZER, T. HOLZ, E. KIRDA, D. BALZAROTTI y C. KRUEGEL, «Abusing social networks for automated user profiling», en *Recent Advances in Intrusion Detection*, Springer, págs. 422–441, 2011.

[9] BERRY, M., *Survey of Text Mining I: Clustering, Classification, and Retrieval*, tomo 1, Springer, 2003.

[10] BING, LIU, «Opinion Lexicon», , 2012, URL `http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html`.

[11] BOLLEN, J., H. MAO y X. ZENG, «Twitter mood predicts the stock market», *Journal of Computational Science*, **2**(1), págs. 1–8, 2011.

[12] BOONE, L. y D. KURTZ, *Contemporary marketing*, South-Western Pub, 2011.

[13] BOUMA, G., «Normalized (pointwise) mutual information in collocation extraction», *Proceedings of GSCL*, págs. 31–40, 2009.

[14] BRIDGE, D., M. GÖKER, L. MCGINTY y B. SMYTH, «Case-based recommender systems», *The Knowledge Engineering Review*, **20**(03), págs. 315–320, 2005.

[15] BRILL, E., «A simple rule-based part of speech tagger», en *Proceedings of the workshop on Speech and Natural Language*, Association for Computational Linguistics, págs. 112–116, 1992.

[16] BROOKE, J., M. TOFILOSKI y M. TABOADA, «Cross-linguistic sentiment analysis: From english to spanish», en *Proceedings of the 7th International Conference on Recent Advances in Natural Language Processing, Borovets, Bulgaria*, págs. 50–54, 2009.

[17] BURT, R. S., «Toward a structural theory of action: network models of social Structure, Perception, and Action.», , 1982.

[18] Cha, M., H. Haddadi, F. Benevenuto y K. Gummadi, «Measuring user influence in twitter: The million follower fallacy», en *4th International AAAI Conference on Weblogs and Social Media (ICWSM)*, págs. 10–17, 2010.

[19] Christakou, C., S. Vrettos y A. Stafylopatis, «A hybrid movie recommender system based on neural networks», *International Journal on Artificial Intelligence Tools*, **16**(05), págs. 771–792, 2007.

[20] Chu, Z., S. Gianvecchio, H. Wang y S. Jajodia, «Who is tweeting on twitter: human, bot, or cyborg?», en *Proceedings of the 26th Annual Computer Security Applications Conference*, ACM, págs. 21–30, 2010.

[21] Dadvar, M., C. Hauff y F. de Jong, «Scope of negation detection in sentiment analysis», , 2011.

[22] De Choudhury, M., «Tie Formation on Twitter: Homophily and Structure of Egocentric Networks», en *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, IEEE, págs. 465–470, 2011.

[23] De Choudhury, M., H. Sundaram, A. John, D. D. Seligmann y A. Kelliher, «"birds of a feather": Does user homophily impact information diffusion in social media», *Arxiv preprint*, 2010.

[24] Debole, F. y F. Sebastiani, «Supervised term weighting for automated text categorization», *Studies in Fuzziness and Soft Computing*, **138**, págs. 81–98, 2004.

[25] Denecke, K., «Using SentiWordNet for multilingual sentiment analysis», en *Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on*, Ieee, págs. 507–512, 2008.

[26] Diakopoulos, N. y D. Shamma, «Characterizing debate performance via aggregated twitter sentiment», en *Proceedings of the 28th international conference on Human factors in computing systems*, ACM, págs. 1195–1198, 2010.

[27] DING, X. y B. LIU, «The utility of linguistic rules in opinion mining», en *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, págs. 811–812, 2007.

[28] DOMINGOS, P. y M. RICHARDSON, «Mining the network value of customers», en *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, págs. 57–66, 2001.

[29] DURAO, F. y P. DOLOG, «A personalized tag-based recommendation in social web systems», *arXiv preprint arXiv:1203.0332*, 2012.

[30] ELLISON, N. *et al.*, «Social network sites: Definition, history, and scholarship», *Journal of Computer-Mediated Communication*, **13**(1), págs. 210–230, 2007.

[31] EVERT, S., «The statistics of word cooccurrences», *Word Pairs and Collocations. Phil. Diss. Institut für maschinelle Sprachverarbeitung. Stuttgart*, 2005.

[32] FALLOWS, D., «The internet and daily life», `http://www.pewinternet.org/~/media/Files/Reports/2004/PIP_Internet_and_Daily_Life.pdf.pdf`, descargado el 23 de Febrero de 2013, 2004.

[33] FELD, S. L., «The focused organization of social ties», *American journal of sociology*, págs. 1015–1035, 1981.

[34] FELDMAN, R. y I. DAGAN, «Knowledge discovery in textual databases (KDT)», en *Proc. 1st Int. Conf. Knowledge Discovery and Data Mining*, págs. 112–117, 1995.

[35] FELDMAN, R. y J. SANGER, *The text mining handbook: advanced approaches in analyzing unstructured data*, Cambridge University Press, 2006.

[36] FELLBAUM, C., «WordNet: An electronic Lexical Database», *Cambridge, MA: MIT Press*, págs. , year=1998, publisher=.

[37] GO, A., R. BHAYANI y L. HUANG, «Twitter-Sentiment», , 2010, URL `http://twittersentiment.appspot.com/`.

[38] GONZÁLEZ-IBÁÑEZ, R., S. MURESAN y N. WACHOLDER, «Identifying sarcasm in Twitter: a closer look», en *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers*, tomo 2, Citeseer, págs. 581–586, 2011.

[39] GOYAL, A., F. BONCHI y L. LAKSHMANAN, «Learning influence probabilities in social networks», en *Proceedings of the third ACM international conference on Web search and data mining*, ACM, págs. 241–250, 2010.

[40] HAN, J. y M. KAMBER, *Data mining: concepts and techniques*, Morgan Kaufmann, 2006.

[41] HIEMSTRA, D., «A probabilistic justification for using tf× idf term weighting in information retrieval», *International Journal on Digital Libraries*, **3**(2), págs. 131–139, 2000.

[42] HOFMANN, T., «Probabilistic latent semantic indexing», en *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, págs. 50–57, 1999.

[43] HOTHO, A., A. NÜRNBERGER y G. PAASS, «A brief survey of text mining», en *LDV Forum-GLDV Journal for Computational Linguistics and Language Technology*, tomo 20, sn, págs. 19–62, 2005.

[44] HUANG, Z., W. CHUNG, T. ONG y H. CHEN, «A graph-based recommender system for digital library», en *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, ACM, págs. 65–73, 2002.

[45] HUBERMAN, B., D. ROMERO y F. WU, «Social networks that matter: Twitter under the microscope», , 2008.

[46] HWEE TAN, A., «Text Mining: The state of the art and the challenges», en *In Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, págs. 65–70, 1999.

[47] JAKOB, N., S. H. WEBER, M. C. MÜLLER y I. GUREVYCH, «Beyond the stars: exploiting free-text user reviews to improve the accuracy of movie recommendations», en *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, ACM, págs. 57–64, 2009.

[48] JANG, H. y H. SHIN, «Effective Use of Linguistic Features for Sentiment Analysis of Korean», , 2011.

[49] JANSEN, B., M. ZHANG, K. SOBEL y A. CHOWDURY, «Twitter power: Tweets as electronic word of mouth», *Journal of the American society for information science and technology*, **60**(11), págs. 2169–2188, 2009.

[50] JANSEN, J., «Online Product Research», `http://www.pewinternet.org/~/media//Files/Reports/2010/PIP%20Online%20Product%20Research%20final.pdf`, descargado el 23 de Febrero de 2013, 2010.

[51] JAVA, A., X. SONG, T. FININ y B. TSENG, «Why we twitter: understanding microblogging usage and communities», en *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, ACM, págs. 56–65, 2007.

[52] JIANG, L., M. YU, M. ZHOU, X. LIU y T. ZHAO, «Target-dependent twitter sentiment classification», *Proc. 49th ACL: HLT*, **1**, págs. 151–160, 2011.

[53] JINDAL, N. y B. LIU, «Identifying comparative sentences in text documents», en *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, págs. 244–251, 2006.

[54] KENNEDY, A. y D. INKPEN, «Sentiment classification of movie reviews using contextual valence shifters», *Computational Intelligence*, **22**(2), págs. 110–125, 2006.

[55] KIM, S.-M. y E. HOVY, «Determining the sentiment of opinions», en *Proceedings of the 20th international conference on Computational Linguistics*, Association for Computational Linguistics, pág. 1367, 2004.

[56] KOTLER, P., H. KARTAJAYA y I. SETIAWAN, *Marketing 3.0*, LID, 2011.

[57] KOULOUMPIS, E., T. WILSON y J. MOORE, «Twitter sentiment analysis: The good the bad and the omg», en *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[58] LAN, M., C. L. TAN, J. SU y Y. LU, «Supervised and traditional term weighting methods for automatic text categorization», *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **31**(4), págs. 721–735, 2009.

[59] LEUNG, C. W., S. C. CHAN y F.-L. CHUNG, «Integrating collaborative filtering and sentiment analysis: A rating inference approach», en *Proceedings of The ECAI 2006 Workshop on Recommender Systems*, Citeseer, págs. 62–66, 2006.

[60] LIDDY, E., «Text mining», *Bulletin of the American Society for Information Science and Technology*, **27**(1), págs. 13–14, 2005.

[61] LIN, C., Y. HE, R. EVERSON y S. RUGER, «Weakly Supervised Joint Sentiment-Topic Detection from Text», *Knowledge and Data Engineering, IEEE Transactions on*, **24**(6), págs. 1134–1145, 2012.

[62] LINDEN, G., B. SMITH y J. YORK, «Amazon. com recommendations: Item-to-item collaborative filtering», *Internet Computing, IEEE*, **7**(1), págs. 76–80, 2003.

[63] LIU, B., «Sentiment Analysis and Subjectivity», en N. Indurkhya y F. J. Damerau (editores), *Handbook of Natural Language Processing, Second Edition*, CRC Press, Taylor and Francis Group, Boca Raton, FL, iSBN 978-1420085921, 2010.

[64] LIU, B., «Sentiment analysis and subjectivity», *Handbook of Natural Language Processing,*, págs. 627–666, 2010.

[65] MAEVE DUGGAN, J. B., «The Demographics of Social Media Users - 2012», `http://www.pewinternet.com/~/media//Files/Reports/2013/PIP_SocialMediaUsers.pdf`, descargado el 23 de Febrero de 2013, 2012.

[66] MALOUF, R. y T. MULLEN, «Taking sides: User classification for informal online political discourse», *Internet Research*, **18**(2), págs. 177–190, 2008.

[67] MANNING, C., P. RAGHAVAN y H. SCHÜTZE, *Introduction to information retrieval*, tomo 1, Cambridge University Press Cambridge, 2008.

[68] MANNING, C. D. y H. SCHÜTZE, *Foundations of statistical natural language processing*, MIT press, 1999.

[69] MCPHERSON, M., L. SMITH-LOVIN y J. M. COOK, «Birds of a feather: Homophily in social networks», *Annual review of sociology*, págs. 415–444, 2001.

[70] MEI, Q., X. LING, M. WONDRA, H. SU y C. ZHAI, «Topic sentiment mixture: modeling facets and opinions in weblogs», en *Proceedings of the 16th international conference on World Wide Web*, ACM, págs. 171–180, 2007.

[71] MICHELSON, M. y S. A. MACSKASSY, «Discovering users' topics of interest on twitter: a first look», en *Proceedings of the fourth workshop on Analytics for noisy unstructured text data*, ACM, págs. 73–80, 2010.

[72] MILLER, G. A., R. BECKWITH, C. FELLBAUM, D. GROSS y K. J. MILLER, «Introduction to wordnet: An on-line lexical database*», *International journal of lexicography*, **3**(4), págs. 235–244, 1990.

[73] MISHNE, G. y N. GLANCE, «Predicting movie sales from blogger sentiment», en *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW 2006)*, tomo 30, págs. 301–304, 2006.

[74] O'CONNOR, B., R. BALASUBRAMANYAN, B. ROUTLEDGE y N. SMITH, «From tweets to polls: Linking text sentiment to public opinion time series»,

en *Proceedings of the International AAAI Conference on Weblogs and Social Media*, págs. 122–129, 2010.

[75] ORTIZ, A., F. CASTILLO y R. GARCÍA, «Análisis de Valoraciones de Usuario de Hoteles con Sentitext: un sistema de análisis de sentimiento independiente del dominio», *Procesamiento de Lenguaje Natural*, **45**(0), págs. 31–39, 2010.

[76] ORTIZ, A., Á. POZO y S. SÁNCHEZ, «Sentitext: sistema de análisis de sentimiento para el español», *Procesamiento de Lenguaje Natural*, **45**(0), págs. 297–298, 2010.

[77] ORTIZ, A. M., C. P. HERNÁNDEZ y R. H. GARCÍA, «Utilización de córpora textuales para la extracción de modificadores contextuales de valencia para tareas de Análisis de Sentimiento», .

[78] PADRÓ, L., M. COLLADO, S. REESE, M. LLOBERES, I. CASTELLÓN *et al.*, «Freeling 2.1: Five years of open-source language processing tools», , 2011.

[79] PANG, B. y L. LEE, *Opinion mining and sentiment analysis*, Now Pub, 2008.

[80] PANG, B., L. LEE y S. VAITHYANATHAN, «Thumbs up?: sentiment classification using machine learning techniques», en *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, Association for Computational Linguistics, págs. 79–86, 2002.

[81] PAZZANI, M. y D. BILLSUS, «Learning and revising user profiles: The identification of interesting web sites», *Machine learning*, **27**(3), págs. 313–331, 1997.

[82] PHELAN, O., K. MCCARTHY y B. SMYTH, «Using twitter to recommend real-time topical news», en *Proceedings of the third ACM conference on Recommender systems*, ACM, págs. 385–388, 2009.

[83] POLANYI, L. y A. ZAENEN, «Contextual valence shifters», *Computing attitude and affect in text: Theory and applications*, págs. 1–10, 2006.

[84] PROVOST, F., B. DALESSANDRO, R. HOOK, X. ZHANG y A. MURRAY, «Audience selection for on-line brand advertising: privacy-friendly social network targeting», en *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, págs. 707–716, 2009.

[85] READ, J., «Using emoticons to reduce dependency in machine learning techniques for sentiment classification», en *Proceedings of the ACL Student Research Workshop*, Association for Computational Linguistics, págs. 43–48, 2005.

[86] ROCCHIO, J. J., «Relevance feedback in information retrieval», , 1971.

[87] RODRIGUEZ, F., «Correlación del sentimiento negativo y positivo de comentarios publicados en Twitter», en *Daena Jornal*, págs. 43–48, 2011.

[88] SALTON, G., *Automatic Text Processing: The Transformation, Analysis, and Retrieval of*, Addison-Wesley, 1989.

[89] SALTON, G. y C. BUCKLEY, «Term-weighting approaches in automatic text retrieval», *Information processing & management*, **24**(5), págs. 513–523, 1988.

[90] SARWAR, B., G. KARYPIS, J. KONSTAN y J. RIEDL, «Analysis of recommendation algorithms for e-commerce», en *Proceedings of the 2nd ACM conference on Electronic commerce*, ACM, págs. 158–167, 2000.

[91] SCHAFER, J., J. KONSTAN y J. RIEDI, «Recommender systems in e-commerce», en *Proceedings of the 1st ACM conference on Electronic commerce*, ACM, págs. 158–166, 1999.

[92] SEBASTIANI, F., «Machine learning in automated text categorization», *ACM computing surveys (CSUR)*, **34**(1), págs. 1–47, 2002.

[93] SHARMA, G. y S. SINGH, «Economic Analysis of Post-harvest Losses in Marketing of Vegetables in Uttarakhand», *Agricultural Economics Research Review*, **24**(2), págs. 309–315, 2011.

[94] STATS, I. W., «Internet World Users by Language», `http://www.internetworldstats.com/stats7.htm`, descargado el 24 de Febrero de 2013, 2010.

[95] STATS, I. W., «Internet usage statistics», `http://www.internetworldstats.com/stats.htm`, descargado el 23 de Febrero de 2013, 2012.

[96] SUBRAMANI, M. y B. RAJAGOPALAN, «Knowledge-sharing and influence in online social networks via viral marketing», *Communications of the ACM*, **46**(12), págs. 300–307, 2003.

[97] SVENDSEN, M., S. HAUGLAND, K. GRØNHAUG y T. HAMMERVOLL, «Marketing strategy and customer involvement in product development», *European Journal of Marketing*, **45**(4), págs. 513–530, 2011.

[98] TABOADA, M., J. BROOKE, M. TOFILOSKI, K. VOLL y M. STEDE, «Lexicon-based methods for sentiment analysis», *Computational Linguistics*, **37**(2), págs. 267–307, 2011.

[99] TAKAMA, Y. y Y. MUTO, «Profile generation from tv watching behavior using sentiment analysis», en *Web Intelligence and Intelligent Agent Technology Workshops, 2007 IEEE/WIC/ACM International Conferences on*, IEEE, págs. 191–194, 2007.

[100] TAN, C., L. LEE, J. TANG, L. JIANG, M. ZHOU y P. LI, «User-level sentiment analysis incorporating social networks», *arXiv preprint arXiv:1109.6018*, 2011.

[101] TAN, S., X. CHENG, Y. WANG y H. XU, «Adapting naive Bayes to domain adaptation for sentiment analysis», *Advances in Information Retrieval*, págs. 337–349, 2009.

[102] TURNEY, P., «Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews», en *Proceedings of the 40th Annual*

Meeting on Association for Computational Linguistics, Association for Computational Linguistics, págs. 417–424, 2002.

[103] TURNEY, P., M. LITTMAN *et al.*, «Measuring praise and criticism: Inference of semantic orientation from association», , 2003.

[104] TURNEY, P. y M. L. LITTMAN, «Unsupervised learning of semantic orientation from a hundred-billion-word corpus», , 2002.

[105] TWITTER, I., «200 million tweets per day», , 2011, URL `http://blog.twitter.com/2011/06/200-million-tweets-per-day.html`.

[106] TWITTER, I., «What Are @Replies and Mentions?», `http://support.twitter.com/articles/14023-what-are-replies-and-mentions#`, descargado el 20 de Enero de 2013, 2013.

[107] VALDIMIR, V. y N. VAPNIK, «The nature of statistical learning theory», , 1995.

[108] WANG, X., L. TANG, H. GAO y H. LIU, «Discovering overlapping groups in social media», en *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, IEEE, págs. 569–578, 2010.

[109] WANG, Y., G. CONG, G. SONG y K. XIE, «Community-based greedy algorithm for mining top-k influential nodes in mobile social networks», en *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, págs. 1039–1048, 2010.

[110] WANG, Z., Q. ZHOU, J. FANG y H. ZHANG, «Incorporating Sentiment Analysis for Improved Tag-based Recommendation», *Audio Engineering*, **7**, pág. 019, 2012.

[111] WIEGAND, M., A. BALAHUR, B. ROTH, D. KLAKOW y A. MONTOYO, «A survey on the role of negation in sentiment analysis», en *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, Association for Computational Linguistics, págs. 60–68, 2010.

[112] YANG, W., J. DIA, H. CHENG y H. LIN, «Mining social networks for targeted advertising», en *System Sciences, 2006. HICSS'06. Proceedings of the 39th Annual Hawaii International Conference on*, tomo 6, IEEE, págs. 137a–137a, 2006.

[113] YI, J., T. NASUKAWA, R. BUNESCU y W. NIBLACK, «Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques», en *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, IEEE, págs. 427–434, 2003.

[114] YOSHII, K., M. GOTO, K. KOMATANI, T. OGATA y H. OKUNO, «An efficient hybrid music recommender system using an incrementally trainable probabilistic generative model», *Audio, Speech, and Language Processing, IEEE Transactions on*, **16**(2), págs. 435–447, 2008.

[115] YU, H. y V. HATZIVASSILOGLOU, «Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences», en *Proceedings of the 2003 conference on Empirical methods in natural language processing*, Association for Computational Linguistics, págs. 129–136, 2003.

[116] ZAVIŠIĆ, S. y Ž. ZAVIŠIĆ, «Social network marketing», en *22. CROMAR Congress*, 2011.

[117] ZHAI, Z., B. LIU, L. ZHANG, H. XU y P. JIA, «Identifying evaluative sentences in online discussions», en *Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.

[118] ZHAO, D. y M. ROSSON, «How and why people Twitter: the role that micro-blogging plays in informal communication at work», en *Proceedings of the ACM 2009 international conference on Supporting group work*, ACM, págs. 243–252, 2009.

[119] Zhao, T., C. Li, Q. Ding y L. Li, «User-sentiment topic model: refining user's topics with sentiment information», en *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, ACM, pág. 10, 2012.

Ficha autobiográfica

Fernando Manuel Rodríguez Aldape

Candidato para el grado de Maestría en Ingeniería de la Información
con orientación en Inteligencia Artificial.

Universidad Autónoma de Nuevo León

Facultad de Ingeniería Mecánica y Eléctrica

Tesis:

Cuantificación del Interés de un usuario en un tema mediante minería de texto y análisis de sentimiento

Nací en la ciudad de Monterrey, Nuevo León, México el año de 1985. Obtuve el grado de Licenciado en Ciencias Computacionales en la Facultad de Ciencias Físico-Matemáticas en el año 2007. En 2009 viajé a la ciudad de Barcelona, España para estudiar un posgrado en Dirección y Gestión de Marketing en la Universitat de Barcelona. Posteriormente, en el año 2010, regresé a mi ciudad y continué mis estudios en la Facultad de Ingeniería Mecánica y Eléctrica en la Maestría en Ingeniería de la Información con Orientación en Inteligencia Artificial a cuyo título actualmente aspiro.