UCLV
Universidad Central
"Marta Abreu" de Las Villas

MFC
Facultad de Matemática
Física y Computación

Laboratorio de Inteligencia Artificial

TRABAJO DE DIPLOMA

Título: Detección de polaridad por tópicos para textos cortos en Español.

Autor: Alejandro Ariel Ramón Hernández

Tutoras: Dra.C María Matilde García Lorenzo

Dra.C Leticia Arco García

Santa Clara

UCLV
Universidad Central
"Marta Abreu" de Las Villas

MFC
Facultad de Matemática
Física y Computación

Artificial Intelligence Laboratory

DIPLOMA THESIS

Title: Polarity detection of opinions by topics in short texts in Spanish.

Author: Alejandro Ariel Ramón Hernández

Thesis Director: Dra.C María Matilde García Lorenzo

Dra.C Leticia Arco García

Santa Clara

DEDICATORIA

A mis padres y mi hermanita linda

AGRADECIMIENTOS

A mi mamá por estar siempre ahí para ayudarme y apoyarme en todo, por hacer posible que llegara hasta aquí. Por dármelo todo sin esperar nada a cambio.

A mi papa porque con él he aprendido mucho y estoy seguro que seguiré aprendiendo. Por cada comida rica que nos ha hecho que me pone a reventar.

A mi hermanita que siempre me espera para ver series juntos en la casa mientras comemos, por soportar mis pesadeces sin ponerse brava y reírse de mis boberías.

A mi tía Yolanda que, aunque no vive cerca de mí, siento como si fuera así, por ser como una segunda mamá y cuidarme y preocuparse por mi desde que soy chiquitico.

A todos mis amigos de la universidad, que entraron siendo desconocidos para mi y saldremos siendo hermanos. En especial a los integrantes de la berrakera que fue ampliando sus filas con los años. Gracias por acompañarme en este viaje que tanto me ha cambiado y que al final se me hizo corto...

A Bety por aguantarme las pesadeces y darme tanto cariño. Por regañarme cada rato para que no pierda el tiempo, que gracias a eso logré terminar la tesis. También a sus padres por acogerme en su casa y hacerme sentir uno más de la familia.

Al profesor Alfredo Simón por su ayuda y consejos en el desarrollo de esta investigación.

A mis tutoras Leticia y María Matilde por ayudarme tanto en este trabajo, guiándome y aconsejándome en todo momento. El resultado de este trabajo es tan suyo como mío.

RESUMEN

Las opiniones son importantes para la toma de decisiones, de ahí la necesidad de analizar variantes para extraer mayor cantidad de información de las mismas. Entre la información relevante que ofrecen, la polaridad que expresa y los tópicos a los que se refieren constituyen aspectos a trabajar. Generalmente la polaridad se calcula a nivel de documento o de oración, por lo que en ocasiones este análisis no brinda información suficiente para la toma de decisiones. En este trabajo se propone un procedimiento para calcular la polaridad de las opiniones agrupadas por tópicos. Para esto un método de agrupamiento del tipo jerárquico aglomerativo se aplica para obtener grupos por tópicos y la polaridad se calcula con el uso del recurso SpanishSentiWordnet. El procedimiento permite conocer el grado de positividad y negatividad expresado sobre los distintos tópicos tratados en un conjunto de opiniones.

ABSTRACT

Opinions are important for decision making, hence the need to analyze variants to extract more information from them. Among the relevant information they offer, the polarity that it expresses and the topics to which they refer constitute aspects to be worked on. Polarity is usually calculated at the document or sentence level, so sometimes this analysis does not provide enough information for decision making. This paper proposes a procedure to calculate the polarity of opinions grouped by topics. For this, a clustering method of the agglomerative hierarchical type is applied to obtain groups by topics and the polarity is calculated with the use of the SpanishSentiWordnet resource. The procedure allows knowing the degree of positivity and negativity expressed on the different topics treated in a set of opinions.

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

INTRODUCCIÓN

Un factor importante a tener en cuenta para tomar una decisión es conocer las experiencias de otras personas en el tema, por lo que las opiniones son de gran importancia en muchos ámbitos de la vida de los seres humanos, e incluso, en el sector empresarial. En la actualidad, el desarrollo de las tecnologías de la informática y las comunicaciones ha propiciado la aparición en internet de nuevos sitios ricos en opiniones. El reto consiste en desarrollar técnicas y herramientas que permitan procesar de forma efectiva estas opiniones con vistas a facilitar la toma de decisiones.

Las opiniones son los estados subjetivos que reflejan los sentimientos y la percepción de una persona sobre un suceso o un objeto (Kaur and Duhan, 2015). Según Bing Liu (Liu, 2015), una opinión es una emoción sobre una entidad o un aspecto de la entidad expresado por un usuario, esta entidad puede ser un producto, persona, evento, organización, o tópico. Una opinión también ha sido definida matemáticamente como una quíntupla (Liu, 2012), donde se incluye el nombre de una entidad, los aspectos de la entidad, los sentimientos de cada uno de los aspectos de la entidad, el titular (usuario) de la opinión, y el momento en que la opinión fue expresada. El sentimiento puede ser positivo, negativo o neutro, o se expresa con distinta fuerza o niveles de intensidad.

Las opiniones tienen gran importancia tanto desde el punto de vista del usuario, así como para una organización política o empresa. Al usuario individual, le permite conocer la experiencia de otros usuarios con el producto o servicio que planea adquirir, y así tomar una decisión más certera. Por otra parte, para una empresa que oferta algún producto o servicio, es de gran importancia conocer las opiniones de los usuarios sobre sus productos para apoyado en estas corregir los posibles errores y mejorar la calidad. Desde el punto de vista de una organización política es importante conocer el estado de opinión de la población ya sea de la organización de forma general o alguna acción que se haya desarrollado.

La gran cantidad de opiniones generadas en un corto período de tiempo, dificulta su análisis manual, por lo que se hace necesario el desarrollo de sistemas de minería de opinión que sean capaces de procesarlas de forma automática o semi-automática.

Existen varias definiciones de minería de opinión o análisis de sentimientos. Según Bing Liu (Liu, 2012), el análisis de sentimientos o la minería de opinión es el estudio computacional de las opiniones, valoraciones, actitudes y emociones expresadas por los usuarios hacia entidades, personas, temas, eventos, productos y sus atributos.

Existen varias definiciones de minería de opinión (Kaur and Duhan, 2015; Vasantharaj *et al.*, 2015; Liu, 2017), unas más generales, otras muy específicas, pero todas tienen como elemento común la aplicación de técnicas del procesamiento del lenguaje natural (Khurana *et al.*, 2017), lingüística computacional (Hutchison, 2013) y minería de textos (Berry, 2004), para la extracción de información subjetiva a partir de contenidos generados por los usuarios (comentarios en blogs, evaluaciones de productos o servicios, y respuestas a una encuesta, etc.) lo cual permitirá contribuir a la toma de decisiones.

Una de las principales tareas de la minería de opinión es la clasificación de la polaridad de la opinión, que consiste en determinar si la opinión es positiva o negativa con respecto a la entidad a la que se refiere, que puede ser una persona, un producto, una temática.

Para la determinación de la polaridad se hace uso de técnicas de aprendizaje supervisado y no supervisado, cada uno con sus ventajas y desventajas. Por una parte, las técnicas de aprendizaje supervisado arrojan muy buenos resultados con la desventaja que son muy dependientes del domino de aplicación y la calidad y tamaño de los conjuntos de entrenamiento. Por otra parte, las soluciones no supervisadas no dependen de casos previamente clasificados, por lo que no son dependientes del dominio de aplicación, pero se basan en recursos externos que actualmente son limitados, además son mayormente dependientes del idioma y en su mayoría son desarrollados para el inglés.

El procesamiento automático de opiniones no es una tarea sencilla. Algunos de los problemas presentes en el tratamiento de las opiniones son: el uso de lenguaje informal, las abreviaturas, los errores ortográficos y tipográficos, el lenguaje irónico y sarcástico, el nivel de conocimiento del lenguaje, el nivel cultural, entre otros. Estos problemas, en comparación con el procesamiento de documentos en otras tareas de la minería de textos, imponen una mayor dificultad a la minería de opiniones.

Debido a estas dificultades es uno de los temas más complejos del procesamiento de lenguaje natural, a su vez es uno de los más demandados por usuarios e instituciones debido a las ventajas que supone procesar automáticamente las opiniones.

La máxima dirección del país ha insistido en reiteradas ocasiones en la necesidad que existe de informatizar la sociedad cubana. Varias empresas, organizaciones, e instituciones en general desean informatizar sus procesos y servicios. Es importante destacar que ha sido un interés de la Asamblea del Poder Popular Provincial (APPP) de Villa Clara analizar automáticamente las opiniones que constantemente llegan de los ciudadanos. Además, la empresa DESOFT Villa Clara trabaja en el desarrollo del portal del ciudadano con el objetivo de informatizar varios procesos y servicios, y, por tanto, repercutir favorablemente en el bienestar ciudadano. Es por ello que resulta interesante desarrollar una herramienta que permita procesar automáticamente las opiniones que se emitan en algún foro de discusión disponible en el portal que se encuentra en desarrollo.

En el laboratorio de Inteligencia Artificial del Centro de Investigaciones de la Informática (CII) de la Universidad Central "Marta Abreu" de Las Villas (UCLV) se han desarrollado varias herramientas que permiten abordar etapas de la minería de textos y, específicamente, de la minería de opinión. De ahí que directivos de la APPP de Villa Clara se hayan acercado al CII buscando colaboración en el desarrollo de herramientas que permitan procesar las opiniones ciudadanas.

Las herramientas PosNeg Opinion (Amores, Arco and Artiles, 2015) y OpinionTopicDetection (Orozco, 2016) desarrolladas por investigadores y estudiantes del CII podrían contribuir al análisis de las opiniones que emiten los villaclareños; sin embargo, ellas tienen algunas limitaciones, lo que motiva el desarrollo de una nueva herramienta basada en las ya existentes y que responda a los intereses de la provincia.

La herramienta PosNeg Opinion detecta la polaridad de las opiniones mostrando excelentes valores de precisión y exactitud, pero solo es capaz de clasificar las opiniones expresadas en una oración o la opinión que expresa el texto en su totalidad.

Por otro lado, OpinionTopicDetection detecta los tópicos de los cuales se emiten opiniones escritas en Inglés, pero no permite calcular la polaridad de las opiniones por tópicos.

Estas herramientas arrojan excelentes resultados, pero solo garantizan algunas etapas de la minería de opinión y no se encuentran integradas, lo que limita el cálculo de la polaridad de opiniones organizadas por tópicos y escritas en español, dando lugar al **planteamiento del problema de investigación** siguiente: Se adolece de una herramienta única que permita procesar opiniones en Español y calcular la polaridad por tópicos.

El **objetivo general de la investigación** consiste en desarrollar una herramienta que permita calcular de manera no supervisada la polaridad de opiniones en Español agrupadas por tópicos. Este se desglosa en los siguientes **objetivos específicos**:

1. Identificar herramientas, módulos, bibliotecas y sistemas en general que permitan realizar alguna etapa de la detección de la polaridad y los tópicos en las opiniones.
2. Diseñar un procedimiento general que establezca las etapas a seguir para la detección de la polaridad de opiniones escritas en Español agrupadas por tópicos.
3. Implementar el procedimiento general para la detección de polaridad de textos cortos en Español agrupadas por tópicos.
4. Evaluar la herramienta implementada a partir de corpus textuales.

La tesis está estructurada en tres capítulos. En el Capítulo 1 se mencionan los elementos esenciales de la minería de opinión y aquellas técnicas de la minería de textos y el procesamiento del lenguaje natural que contribuyen a la minería de opinión, profundizando en aquellos métodos y herramientas que permiten determinar la polaridad de las opiniones, el procesamiento textual, el agrupamiento textual y la detección de tópicos. En el Capítulo 2 se propone un esquema general, que consta de cinco etapas, para detectar la polaridad de las opiniones agrupadas por tópicos. En el tercer capítulo se presenta el software PosNegTopicOpinion que implementa el procedimiento propuesto y su evaluación. Este documento culmina con las conclusiones, recomendaciones y referencias bibliográficas.

CAPÍTULO 1. ACERCA DE LA MINERÍA DE OPINIÓN

En este capítulo se abordan aspectos de la minería de opinión y se profundiza en tareas específicas como el cálculo de la polaridad y el resumen en las opiniones describiendo las principales técnicas para cada tarea. También las tareas básicas de la minería de textos son tratadas, en particular aquellas relevantes para la minería de opinión, tales como los diferentes enfoques para la detección de tópicos en las variantes supervisadas y no supervisadas. Finalmente, se analizan los diferentes recursos y herramientas que contribuyen a la minería de textos y de opinión.

1.1 Minería de opinión

La minería de opinión o el análisis de sentimiento es un área de investigación que se encarga de explorar y descubrir la información subjetiva generada por el usuario. Se ha definido como "la tarea de detectar, extraer y resumir opiniones, polaridades y/o emociones, basadas en la presencia o ausencia de rasgos de sentimiento" (Moens, Li and Chua, 2014). En (Vasantharaj *et al.*, 2015) se definió ésta como: Dado un conjunto de documentos textuales *D* que contiene las opiniones (o sentimientos) en relación con un objeto, la minería de opinión pretende extraer atributos y elementos de lo que se ha comentado en cada documento *d* que pertenece a *D* y conocer si los comentarios son o no positivos, negativos o neutros.

En (Kaur and Duhan, 2015) definen la minería de opinión como la técnica para extraer la información subjetiva del texto y determinar la polaridad global de la opinión. Sin embargo, existen oraciones objetivas que expresan opinión y oraciones subjetivas que no expresan ninguna opinión. Por otro lado, puede ser interesante en la toma de decisiones calcular la polaridad local, por aspectos o tópicos. Por lo tanto, esta definición ofrece una visión limitada del análisis de sentimiento.

Los trabajos desarrollados en esta área se inspiran en el hecho que las opiniones de usuarios no expertos pueden servir como complemento de puntos de vistas publicados en diferentes medios; y las valoraciones sobre productos y servicios pueden tener gran impacto económico para los consumidores y organizaciones.

Existen tres niveles fundamentales para desarrollar el análisis de sentimiento (Zhang and Liu, 2014), ellos son:

- Nivel de documento: considera que cada documento expresa una opinión negativa o positiva de forma general; no da detalles de las preferencias de los usuarios, solo se tiene una valoración global de la polaridad de la opinión, pero no se sabe sobre cuáles aspectos específicos se tiene una opinión positiva o negativa
- Nivel de oración: asume que las oraciones expresan opiniones positivas y negativas. Es un nivel más detallado, pero solo tiene en cuenta las palabras de opinión. La cantidad de palabras positivas y negativas son contadas a partir de las oraciones identificadas. Si predominan las palabras positivas la opinión es positiva, de lo contrario son las negativas y en caso que sean iguales es neutral
- Nivel de aspecto: considera que las opiniones son detalladas y expresan sentimientos sobre diferentes aspectos de entidades y de las entidades en sí mismas. Este nivel está muy relacionado con la detección de tópicos, debido a que los aspectos pueden representar tópicos en el texto. Se define a partir de los atributos de las entidades, para los que se identifican opiniones negativas o positivas. Si el texto está mal escrito gramaticalmente puede que no de buenos resultados. Sin embargo, es el modelo que brinda mayor información de los tres mencionados (Sharma and Chitre, 2014).

  Para obtener un análisis de opinión más preciso se recomienda el nivel de aspecto y para ello se ha propuesto combinar tres subtareas (Zhang and Liu, 2014): identificar y extraer entidades en textos, identificar y extraer aspectos de entidades y posteriormente determinar polaridades del sentimiento expresado en entidades y aspectos de entidades.

  De forma general, una opinión se representa por la combinación de cinco componentes: el nombre de una entidad, un aspecto de la entidad, la orientación de la opinión sobre el aspecto de la entidad, el propietario de la opinión, y el tiempo donde la opinión es expresada por el propietario de la opinión.

  La orientación de la opinión se conoce como orientación del sentimiento, polaridad de la opinión u orientación semántica y puede clasificarse en positiva, negativa o neutral, o expresarse con diferentes niveles de intensidad (Liu, 2010). Un término tiene polaridad cuando porta información subjetiva positiva o negativa.

1.1.1 Cálculo de la polaridad de las opiniones

La d**etección de la polaridad de una opinión consiste en** determinar si una opinión es **positiva o negativa**. Más allá de una polaridad básica, también se puede querer obtener un **valor numérico** dentro de un rango determinado, que de una determinada forma trate de obtener una clasificación objetiva asociada a una determinada opinión.

Se dice que un término tiene polaridad u orientación cuando porta información bien sea positiva o negativa. En este sentido, las colocaciones pueden explotarse puesto que determinados términos pueden adquirir o cambiar su polaridad dependiendo de si forma parte o no de una colocación. Por ejemplo, el adjetivo *alto*, un término que a priori no tiene polaridad de ningún tipo, pero que adquiere polaridad negativa al formar parte de la colocación *precio alto*. El mismo adjetivo en colocación con *valor* tiene, por el contrario, polaridad positiva. El problema está en cómo asignar automáticamente la polaridad.

Los cálculos de polaridad se pueden estructurar en varias fases (Kim *et al.*, 2004) entre ellas:

1. Clasificar la opinión: consiste en determinar la polaridad de la opinión, es decir, si la opinión es negativa o positiva.
2. Determinar la fuerza de la opinión: permite expresar en qué medida la opinión es positiva o negativa.
3. Determinar la fuente de la opinión: consiste en identificar si la fuente de la opinión fue una persona o una institución, requiere frecuentemente resolución de anáforas.
4. Determinar el objetivo de la opinión: se encarga de determinar de quién se habla en la opinión, con quién se está de acuerdo o no.
5. Resumir las opiniones y/o visualizar gráficamente los resultados: consiste en expresar y presentar los valores de polaridad calculados ya sea agregando votos (índice de 1-5, estrellas), sobresaltando algunas opiniones para representar acuerdo/desacuerdo, entre otros.

Actualmente existen dos aproximaciones principales para resolver automáticamente la polaridad de un texto (Kaur and Duhan, 2015):

1. Aprendizaje supervisado: Se construye un clasificador a partir de un conjunto de entrenamiento formado por una colección de textos etiquetados, donde se expresa una opinión favorable o desfavorable. Se trata de un entrenamiento supervisado en el que el clasificador aprenderá a reconocer en base a los ejemplos que se le presenten.

2. Orientación semántica: Se emplean diccionarios donde cada palabra se encuentra etiquetada con su orientación semántica, esto permite medir en qué grado esa palabra es positiva o negativa. El sentimiento del texto se obtiene agregando los valores de los términos del texto que aparezcan en alguno de los diccionarios. Existen dos formas de generar estos diccionarios: crearlos de forma manual o construirlos de forma semiautomática, utilizando palabras semilla. En este último caso, el diccionario se va ampliando en un proceso iterativo atendiendo a relaciones de proximidad física. Se parte de un grupo reducido de palabras de polaridad extrema, como "felicidad" o "asesino" y, posteriormente, en cada iteración palabras cercanas a términos positivos o negativos pasarán a considerarse también positivas o negativas, respectivamente.

   Los clasificadores obtenidos a partir de la primera alternativa se caracterizan por conseguir un buen rendimiento base para el dominio en el que son entrenados. Sin embargo, presentan complicaciones para mejorar su precisión. Además, las soluciones desarrolladas empeoran drásticamente su rendimiento cuando se utilizan para analizar textos de un dominio diferente al del corpus con el que se entrenaron. Es por ello que en la actualidad se trabaja en el aprendizaje continuo (*lifelong learning*) para detectar la polaridad de las opiniones en diversos dominios (Chen and Liu, 2016).

   La segunda alternativa permite una mejor adaptación a los diferentes dominios y contempla más aspectos del texto. El principal inconveniente radica en que los recursos deben construirse cada vez que se quiere aplicar a un nuevo idioma, mientras que empleando aprendizaje automatizado es suficiente con obtener un conjunto de entrenamiento y clasificar los documentos.

   A manera de resumen, las técnicas no supervisadas tienen en cuenta para determinar la positividad y la negatividad la presencia de palabras disparadoras de sentimientos con orientaciones conocidas obtenidas de diccionarios o corpus. En cambio, en las supervisadas, los rasgos extraídos del texto y el método de aprendizaje determinan cuando la opinión pertenece a la clase positiva o negativa.

1.1.2 Resumen de las opiniones

Una tarea muy importante que contribuye a la efectiva toma de decisiones es el resumen de las opiniones y la visualización gráfica de los resultados. El resumen de textos se ha estudiado

ampliamente en el procesamiento del lenguaje natural (Das and Martins, 2007). Sin embargo, un resumen de opiniones es muy diferente del resumen de un documento tradicional, ya que al resumir opiniones generalmente es necesario centrarse en las entidades, los aspectos y los sentimientos acerca de ellos. Además, es necesario incluir elementos cuantitativos, que son la esencia del resumen basado en los aspectos de la opinión (Raut and Londhe, 2014).

El resumen puede estar en una forma estructurada o no estructurada, como un documento de texto corto, los componentes claves de un resumen deben incluir opiniones sobre diferentes entidades y sus aspectos y también deben tener un punto de vista cuantitativo (Liu, 2015).

La perspectiva cuantitativa es especialmente importante, ya que se desprenden análisis diferentes si se conoce que el 20% de los clientes tienen una opinión positiva que si fuera el 90% opinando de manera positiva.

El proceso de resumen de la opinión se centra principalmente en los dos enfoques siguientes: basado en características (Liu *et al.*, 2012; Bahrainian and Dengel, 2013; Ranade *et al.*, 2013) cuando implica hallazgo de términos frecuentes (características) que aparecieron en muchos comentarios, y cuando se presenta mediante la selección de frases que contienen información característica particular, por ejemplo, utilizando análisis semántico latente (*Latent Semantic Analysis*; LSA) (Liu *et al.*, 2012; Steinberger, 2013).

El resumen basado en los aspectos de la opinión tiene dos características principales. En primer lugar, se capta la esencia de las opiniones: objetivos de opinión (entidades y sus aspectos) y sentimientos acerca de ellos. En segundo lugar, es cuantitativa, lo que significa que da el número o porcentaje de personas que tienen opiniones positivas o negativas sobre las entidades y aspectos.

Esta forma de resumen estructurado se ha adoptado por investigadores para resumir críticas de películas (Zhuang, Jing and Zhu, 2006), para resumir el texto de la opinión (Hu *et al.*, 2010), y para resumir evaluaciones de servicios (Blair-Goldensohn *et al.*, 2008). Sin embargo, hay que señalar que el resumen basado en el aspecto no tiene por qué ser estructurado; pudiera ser en forma de un documento de texto basado en la misma idea.

Algunas mejoras realizadas a las propuestas iniciales de la creación de resúmenes de las opiniones basándose en aspectos fueron publicadas en (Lerman, Blair-Goldensohn and

McDonald, 2009; Ganesan, Zhai and Han, 2010; Hu *et al.*, 2010; Lu *et al.*, 2010; Nishikawa *et al.*, 2010a, 2010b; Tata and Di Eugenio, 2010; Yatani *et al.*, 2011; Carenini, Cheung and Pauls, 2013).

Varios investigadores también estudiaron el problema de resumir las opiniones mediante la búsqueda de puntos de vista contrastantes (Kim and Zhai, 2009; Lerman and McDonald, 2009; Paul, Zhai and Girju, 2010; Park, Lee and Song, 2011). Otros autores han estudiado el resumen de opiniones de la manera tradicional, es decir, la idea es ofrecer un resumen del texto sin especificar los aspectos (o temas) y sentimientos acerca de ellos (Beineke *et al.*, 2003; Seki *et al.*, 2006; Wang and Liu, 2011).

Una desventaja de los resúmenes tradicionales es que no consideran o consideran muy poco las entidades, aspectos, y sentimientos acerca de ellos, por lo tanto, pueden seleccionar frases que no estén relacionadas con los sentimientos. Además, no ofrecen un punto de vista cuantitativo, que a menudo es importante en la práctica.

La minería de opinión constituye un caso particular de la minería de texto, por esta razón algunos aspectos de la minería de texto deberán ser trabajados. A continuación, se detallan los elementos relativos a la minería de textos útiles para esta investigación.

1.2 Minería de textos y procesamiento textual

La *minería de textos* (Aggarwal and Zhai, 2012) es el proceso de analizar colecciones de materiales de texto con el objeto de descubrir los temas y conceptos claves y descubrir las relaciones ocultas y las tendencias existentes sin necesidad de conocer las palabras o los términos exactos que los autores utilizaron para expresar dichos conceptos.

La minería de textos y la acción de recuperar información son conceptos que a veces se confunden, aunque son bastante diferentes. Una recuperación precisa de la información y su almacenamiento supone un reto importante, pero la extracción y administración de contenido de calidad, de terminología y de las relaciones contenidas en la información son procesos cruciales y determinantes. En otras palabras, la minería de textos es el proceso encargado del descubrimiento de información que no existía explícitamente en ningún texto de la colección,

pero que surge de relacionar el contenido de varios de ellos. Para ello, la minería de textos comprende tres actividades fundamentales (Berry, 2004):

1. Recuperación de la información: seleccionar los textos pertinentes.
2. Extracción de la información: aplicar técnicas de procesamiento del lenguaje natural para identificar hechos, acontecimientos, datos clave, relaciones entre ellos, etc.
3. Minar los datos: encontrar asociaciones entre los datos claves previamente extraídos.

Estas actividades se dividen en tres etapas fundamentales:

- **Preprocesamiento:** los textos se transforman en algún tipo de representación estructurada o semiestructurada que facilite su posterior análisis.
- **Descubrimiento**: las representaciones internas se analizan con el objetivo de descubrir en ellas algunos patrones interesantes o nueva información.
- **Visualización**: los usuarios pueden observar y explorar los resultados.

Cada una de estas etapas presentan retos que pueden superarse desde distintos enfoques lo que depende del dominio de aplicación y el nivel de información que se desea extraer de los textos. De manera general, es necesario aplicar una combinación de técnicas que permita extraer de forma eficiente la información que no aparece en el texto de forma explícita y pueda resultar interesante. Para cumplir esta meta es necesario aplicar técnicas lingüísticas con el objetivo de llevar los datos a un formato estructurado más fácil de procesar.

Estrechamente relacionado con las técnicas de la minería de textos, se encuentran las tareas del Procesamiento del Lenguaje Natural (*Natural Language Processing*; NLP) es un conjunto de procesos informáticos que se encargan de analizar e interpretar textos. Tiene su origen en disciplinas variadas, principalmente del área lingüística, la ciencia de la computación, así como pequeños aportes de la psicología y la lógica.

Tradicionalmente, el NLP ha sido visto como un proceso compuesto por un número de pasos contrastando las diferentes teorías entre la sintaxis, semántica y pragmática. Analizar al texto primero en términos de sintaxis provee de un orden y estructura que hacen posible el análisis en términos de semántica; y luego el paso de análisis pragmático se intenta explicar la

elección de determinadas formas de realizar el enunciado en función de los factores contextuales. Sin embargo, esta visión tripartita compuesta por sintaxis, semántica y pragmática, solo sirve como un punto de partida si se quiere analizar texto en lenguaje natural real, como es el caso del que se podría extraer en redes sociales, en donde lo contextual juega un papel crucial, y en donde se pueden presentar desafíos difíciles de sortear, como, por ejemplo, sarcasmo.

Dentro de las tareas del NLP, en esta investigación resultan de interés aquellas que realizan análisis sintáctico, semántico y de discurso. Se destacan dentro de las sintácticas las dedicadas al etiquetamiento de los términos en partes de la oración (*Part-of-speech tagging*; etiquetamiento POS) y la delimitación en oraciones. Especial atención tienen las siguientes tareas para el análisis semántico: extracción de relaciones, análisis de sentimiento, reconocimiento y segmentación en tópicos. Finalmente, dentro del análisis de discurso es imprescindible mencionar el resumen automático y la resolución de co-referencias.

1.3 Detección de tópicos

La detección de tópicos inicialmente fue declarada como una tarea dependiente de la segmentación, debido a que la entrada de los algoritmos de detección estaba representada por segmentos (Allan, 2002). Sin embargo, varias han sido las propuestas que utilizan como entrada el corpus textual sin segmentar, es decir, aplican técnicas que extraen los términos de los documentos, los agrupan y estos grupos representan los tópicos.

De esta forma se define a la detección de tópicos como: “la tarea que automáticamente encuentra nuevos tópicos en datos textuales” (Huang *et al.*, 2013) o “el proceso de agrupar documentos con tópicos similares en el mismo grupo*”* (Ye *et al.*, 2006).

Debido a la diversidad de fuentes de información surgidas en los últimos años y la gran variedad de temas que se tratan en éstas, hay un auge en el surgimiento de algoritmos para la detección de tópicos. Las tendencias en las implementaciones se pueden dividir en supervisadas y no supervisadas, de acuerdo a si se posee o no un conjunto de documentos de entrada previamente asignados a tópicos. Si se tiene el conocimiento sobre ese conjunto entonces la detección es supervisada, generalmente se requieren expertos del dominio para

decidir cuándo un documento pertenece a un tópico predefinido; se parte de tópicos predefinidos y se entrenan algoritmos para identificar tópicos predefinidos en documentos textuales, luego se predicen los tópicos para nuevos documentos utilizando la experiencia adquirida en el entrenamiento. Por otra parte, para el enfoque no supervisado no se cuenta con ese conjunto inicial de documentos de entrenamiento, sino que los algoritmos propuestos descubren tópicos sin involucrar expertos del dominio (Dong, Hui and He, 2006).

El enfoque supervisado tiene la desventaja que se requiere gran esfuerzo para entrenar los clasificadores. El enfoque no supervisado detecta los tópicos basándose en las similitudes entre sus unidades textuales.

Otra clasificación interesante ubica los métodos de detección de tópicos textuales en tres clases (Petkos, Aiello and Skraba, 2014):

- Métodos documento-pivote: agrupan a documentos individuales de acuerdo a su similitud.
- Métodos rasgo-pivote: agrupan términos de acuerdo a sus patrones de coocurrencia.
- Modelos de tópicos probabilísticos: tratan el problema de la detección de tópicos como un problema de inferencia probabilístico.

Los métodos que pertenecen a cada una de estas clases representan los tópicos de forma diferente. Un documento-pivote representa un tópico con un conjunto de documentos relevantes; un método de rasgo-pivote representa tópicos con un conjunto de términos y un modelo de tópico probabilístico representa un tópico por distribuciones de términos.

Es difícil concluir cuál clase produce los mejores resultados. La similitud de pares de documentos puede ser dominada fácilmente por características ruidosas y por ello los objetos pueden ser incorrectamente agrupados al usar métodos documento-pivote. Los enfoques probabilísticos producen buenos resultados; sin embargo, son computacionalmente muy costosos (Titov and Mcdonald, 2008).

Los enfoques de documento-pivote pueden calcular típicamente una medida de similitud entre un par de documentos o entre un documento y un grupo. En el primer caso, si la similitud entre el documento entrante y el mejor documento coincidente que se encuentra en la colección está por encima de un umbral, entonces el documento entrante se adiciona al mismo grupo como el mejor documento coincidente. De lo contrario, se genera un nuevo grupo. Las diferencias entre los distintos métodos con este enfoque radican en la forma de

calcular la similitud entre los elementos a agrupar y la forma de representación de los documentos; estos métodos pueden también incluir alguna etapa de post procesamiento para refinar los resultados (Petkos, Aiello and Skraba, 2014).

Los métodos rasgos-pivote intentan agrupar términos de acuerdo a sus patrones de coocurrencia. Generalmente, primero se seleccionan los términos a agrupar y se calculan los patrones de coocurrencia. En el segundo paso algunos calculan similitudes entre términos y son generalmente usadas en conjunción con algún procedimiento de agrupamiento. En la literatura se presenta una gran variedad de formas de seleccionar un conjunto de términos para ser agrupados, para calcular la similitud entre términos y ejecutar el agrupamiento (Petkos, Aiello and Skraba, 2014). La mayoría de los métodos rasgos-pivote, independientemente de emplear un mecanismo de selección de términos, examinan los patrones de coocurrencia entre los pares de términos. En la práctica, dependiendo del algoritmo de agrupamiento que se emplee, existe la posibilidad de agrupar términos incorrectamente, especialmente cuando son términos comunes que pueden quedar enlazados a una gran cantidad de tópicos.

Los modelos de tópicos probabilísticos representan la distribución conjunta de tópicos y términos usando un modelo generativo probabilístico, el cual consiste de un conjunto de variables latentes que representan tópicos, términos, hiperparámetros, etc.

1.4 Agrupamiento de documentos

Para realizar análisis de grupos se ha propuesto una gran variedad de algoritmos de agrupamiento (Pascual, Pla and Sánchez, 2007). Estos pueden clasificarse de diversas formas (Magdaleno Guevara, 2015) atendiendo a: tipo de los datos de entrada del algoritmo, criterios para definir la similitud entre los objetos, conceptos en los cuales se basa el análisis y forma de representación de los datos, entre otros. En la Figura 1 se muestra una taxonomía de algoritmos de agrupamiento donde se distinguen dos tipos: los que forman particiones y los jerárquicos.

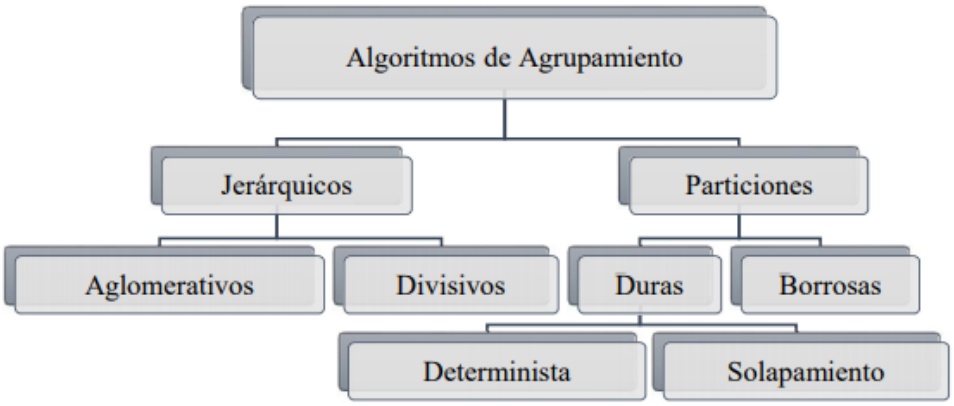

Figura 1: Taxonomía de algoritmos de agrupamiento

Los métodos que forman particiones tienen como objetivo encontrar la mejor partición de los datos en k grupos ($k \in \mathbb{N}$, $k > 0$) basada en una medida de similitud dada y conservar el espacio de particiones posibles en k subconjuntos solamente. Estos algoritmos tienen como desventaja que hay que conocer previamente el número de grupos a formar, o aplicar el agrupamiento para distintos valores de k hasta encontrar una cantidad de grupos óptima, debido a esto no es eficiente aplicar este tipo de métodos si no se tiene información previa de la estructura del conjunto de datos. Uno de los algoritmos más usados dentro de esta clasificación es el *K-Means* (Zhang, 2012).

Por otra parte, los algoritmos jerárquicos hacen una descomposición jerárquica de los objetos. Dentro de ellos, los aglomerativos consideran que cada objeto constituye un grupo, por tanto, inicialmente existen tantos grupos como objetos tiene la colección, y sucesivamente los une, hasta que todos los objetos formen un único grupo, generalmente con una medida de similitud, en este grupo de algoritmos se encuentra el Agrupamiento Jerárquico Aglomerativo (*Hierarchical Agglomerative Clustering*; HAC) (Zhu, 2010). Los divisivos consideran inicialmente que existe un único grupo al cual pertenecen todos los objetos y sucesivamente dividen los grupos, hasta que cada grupo contenga un único objeto.

La construcción de la jerarquía se puede detener por criterios automáticos o del usuario, atendiendo a una cantidad específica de grupos o a un umbral de similitud. Esta última variante permite que estos tipos de métodos de agrupamiento puedan ser aplicados sin tener

ningún conocimiento previo del conjunto de datos, además, el umbral puede calcularse dinámicamente, lo cual favorece ajustarse mejor al dominio.

HAC resulta adecuado como método de agrupamiento a emplear en el presente trabajo en su variante de umbral para la construcción de la jerarquía. Esta selección se basa en el hecho de que no se tiene ningún conocimiento previo de los conjuntos de datos, por lo tanto, no se puede especificar un número específico de grupos a formar. Además, este método no es sensible al orden en que aparecen los datos lo que disminuye la posibilidad de generar grupos ilógicos.

Otro factor importante a tener en cuenta que influye en la calidad del agrupamiento es la elección de la medida de similitud. Una medida de similitud o función de similitud es una función real que cuantifica la similitud entre dos objetos. Toda función de similitud (o de distancia) calcula un valor que permite obtener una medida de proximidad o distancia entre dos documentos dados. Algunas de las similitudes y distancias más utilizadas para comparar objetos son: la distancia Euclidiana, distancia Minkowski, Correlación de Pearson, entre otras (Demey *et al.*, 2015). Entre las funciones más conocidas para el trabajo con colecciones textuales se encuentran: Coseno, Jaccard y Dice (Blair-Goldensohn *et al.*, 2008).

La elección de una medida de similitud adecuada no es trivial y el desempeño de muchos algoritmos depende de la selección de una buena función que se acoja a los datos (KumarPatidar, Agrawal and Mishra, 2012). Para encontrar los agrupamientos naturales, la noción de similitud debe ser adaptada al problema particular; es por ello que actualmente se trabaja en la obtención de medidas que trabajen sobre tipos de datos específicos.

La similitud entre documentos se puede calcular teniendo en cuenta solamente la información sintáctica del texto, sin embargo, en ocasiones esta información no es suficiente para establecer una relación lógica entre estos. Entonces se hace necesario agregar información semántica que permita calcular la similitud de forma más efectiva.

En el caso de documentos cortos estos problemas se acentúan debido a la baja frecuencia de términos. Existen varios trabajos enfocados en encontrar medidas de similitud que tengan en cuenta la información semántica (Mihalcea, Corley and Strapparava, 2006), en estos trabajos se muestran resultados que superan los obtenidos a partir del análisis sintáctico clásico.

La medida de similitud entre los textos T_1 y T_2, expresada en la Ecuación (1), donde *w* es un término perteneciente a T_1 o T_2 , es recomendada por (Mihalcea, Corley and Strapparava, 2006) para el caso de textos cortos donde el cálculo de la frecuencia de aparición de palabras no es suficiente para la conformación de grupos que brinden información de calidad. Esta medida tiene en cuenta el significado de las palabras de los documentos y a partir de este significado establece relaciones de similitud entre ellas lo que se traduce en una mayor comprensión del texto y, por tanto, en un valor de similitud entre documentos más acertado. Para el cálculo de la similitud entre palabras se hace uso de la red semántica WordNet.

$$\text{sim}(T_1, T_2) = \frac{1}{2}\left(\frac{\sum_{w\epsilon\{T_1\}}\left(maxSim(w, T_2) * idf(w)\right)}{\sum_{w\epsilon\{T_1\}} idf(w)} + \frac{\sum_{w\epsilon\{T_2\}}\left(maxSim(w, T_1) * idf(w)\right)}{\sum_{w\epsilon\{T_2\}} idf(w)}\right) \quad (1)$$

1.5 Herramientas y recursos que contribuyen a la minería de textos y de opinión.

De acuerdo con (Esuli 2008), uno de los factores principales para cualquier trabajo relacionado con la minería de opinión es la necesidad de identificar cuáles elementos del lenguaje contribuyen a expresar opinión en un texto. Esa identificación puede ser realizada a través de recursos léxicos que listan propiedades relevantes relacionadas a los ítems de opinión.

Según (Pasqualotti 2008), un recurso léxico es una estructura sistemática que explícitamente expresa determinadas características o significados asociados a las palabras. Cada término de un recurso léxico consiste de una forma ortográfica y fenológica con un formato de representación del significado.

En el contexto de la minería de opinión, los recursos léxicos se construyen o adaptan con la intención de aplicarlos en investigaciones para la detección de la subjetividad en textos y la clasificación de sentimiento. Algunos recursos son compuestos por un conjunto de términos, relacionados a las categorías, por ejemplo, positiva o negativa. Posiblemente los términos

que no están relacionados a una de estas dos categorías pueden ser considerados como objetivos y no contribuir a la orientación global del texto (Esuli 2008).

En el presente trabajo se requieren utilizar varios recursos léxicos para poder desarrollar la minería de opinión por tópicos. A continuación, se describen estos recursos.

WordNet[1]

WordNet es una extensa base de datos léxica de la lengua inglesa, desarrollada por George A. Miller, a partir de la combinación de informaciones lexicográficas (relaciones léxicas y semánticas usadas para representar la organización del conocimiento léxico) y recursos computacionales (Fellbaum, 1998).

En (Miller, 1995) se define el vocabulario de un lenguaje con un conjunto de pares palabra-sentido a partir de un conjunto de significados. En WordNet un sentido se representa por un conjunto de uno o más sinónimos. La base posee más de 118000 palabras diferentes y más de 90000 sentidos.

WordNet está compuesto de sustantivos, verbos, adjetivos y adverbios agrupados en conjuntos de sinónimos (*synsets*), donde cada uno expresa un concepto distinto. Estos son organizados en un conjunto de lexicógrafos por categoría sintáctica y por otros criterios de organización. Los adverbios son mantenidos en apenas un archivo, mientras los sustantivos y verbos son agrupados de acuerdo con la semántica. Los adjetivos son divididos en adjetivos descriptivos y relacionales (Miller *et al.*, 1990).

Para el desarrollo de este trabajo se utilizó la versión 1.6 de WordNet para español.

SentiWordNet[2]

Otro léxico existente, desarrollado por (Esuli, 2008) explícitamente para colaborar en aplicaciones para la minería y clasificación de opiniones, es el SentiWordNet (SWN). Este recurso léxico es resultado de anotaciones automatizadas en todos los synsets del WordNet, con grado de positividad, negatividad y neutralidad. A cada synset se le asocian tres valores numéricos, Pos(*s*), Neg(*s*) y Obj(*s*) que indican cuan positivo, negativo u objetivo (neutro)

[1] https://wordnet.princeton.edu/
[2] http://sentiwordnet.isti.cnr.it/

son los términos contenidos en el synset *s*. Cada uno de los tres valores varían en el intervalo [0.0, 1.0] y la suma de ellos es 1.0 para cada synset.

Para construir los SWN1.0 (Esuli and Sebastiani, 2006b) y 1.1 (Esuli and Sebastiani, 2007) fueron empleados algoritmos de aprendizaje supervisado y semi-supervisados. Para obtener las versiones 2.0 (Esuli and Sebastiani, 2009) y 3.0 (Baccianella, Esuli and Sebastiani, 2010) los resultados de los algoritmos semi-supervisados fueron adoptados como una etapa intermedia en el proceso de etiquetamiento, y posteriormente se siguió un proceso iterativo de refinamiento de los resultados hasta lograr convergencia. La ambigüedad en el SWN 1.0 es tratada por la representación semántica de los synsets del WordNet. Para las demás versiones fueron adoptados recursos diseñados expresamente para manejar la ambigüedad de los términos. Así, el SWN 2.0 adoptó el eXtended WordNet[3] y el SWN 3.0 adoptó el Princeton WordNet Gloss Corpus[4], que asume ser más preciso que el eXtended WordNet.

El método para desarrollar el SentiWordNet es derivado de los trabajos de (Esuli and Sebastiani, 2005, 2006a), basado en el análisis cuantitativo de términos asociados a synsets y en el uso de la representación vectorial de términos resultantes para la clasificación semi-supervisada de synsets. El trío de puntuación en SWN se deriva de la combinación de los resultados producidos por un conjunto de ocho clasificadores ternarios, con similares niveles de precisión; sin embargo, con diferentes comportamientos de clasificación. Un clasificador difiere del otro teniendo en cuenta el conjunto de formación de partida y el aprendizaje adoptado para este conjunto, produciendo así distintas clasificaciones resultantes de cada synset del WordNet.

La puntuación en el SWN es dada por la proporción de clasificadores que señaló la correspondiente etiqueta al synset. Si todos los clasificadores ternarios resultan en atribuir la misma etiqueta a un synset, la etiqueta tendrá la máxima puntuación para el synset, de lo contrario tendrá puntuación proporcional al número de clasificadores que señaló (Esuli, 2008; Baccianella, Esuli and Sebastiani, 2010). El SWN 3.0 posee versión web para consulta online[5]. La base de datos del SWN también está disponible en un archivo de datos (.txt). La **¡Error! No se encuentra el origen de la referencia.** muestra algunos ejemplos de registros e

[3] http://www.hlt.utdallas.edu/~xwn/downloads.html
[4] http://wordnet.princeton.edu/glosstag.shtml
[5] http://sentiwordnet.isti.cnr.it/

n SentiWordNet 3.0. El POS-ID identifica el synset, los valores PosScore y NegScore corresponden a la positividad y negatividad señalados por el SentiWordNet para el determinado synset. El valor de objetividad es dado por: Obj(s)=1-(Pos(s)+Neg(s)). SynsetTerms corresponde a los términos separados por espacio, pertenecientes al synset, con la clase gramatical y número correspondiente al sentido. Gloss describe el sentido del término.

POS	ID	PosScore	NegScore	SynsetTerms	Gloss
A	00016247	0.125	0.500	super abundant #1	most excessively abundant
A	00122245	0.375	0.125	prevenient #1 anticipatory #1	in anticipation
R	00124611	0.000	0.125	legally#2	in a legal manner; "he acted legally"

Tabla 1: Ejemplos de registros del SentiWordNet 3.0

SentiWordNet cubre todos los synsets de WordNet. Aunque este es quizás uno de sus mayores aportes, es también una de sus mayores debilidades, puesto que, en muchos casos, conceptos que no estén bien relacionados con el conjunto inicial de synsets etiquetados manualmente no obtendrán puntuaciones muy acordes al etiquetado esperado. Un ejemplo de esto es que el concepto representativo de "tumor" o "cáncer" obtiene una puntuación muy baja (0.125) en la categoría Neg.

El SentiWordNet se construyó semiautomáticamente; por tanto, todos los resultados no fueron validados manualmente, por lo que, algunas clasificaciones pueden ser incorrectas. Por ejemplo, el synset#1 del sustantivo "flu"({influenza, flu, grippe} - (*an acute febrile highly contagious viral disease*)) se clasificó como Positivo=0.75, Negativo=0.0, Objetivo=0.25, a pesar de tener varias palabras negativas en su glosario.

TreeTagger[6]

El TreeTagger es una herramienta para anotar textos con información de part-of-speech y lema, desarrollado dentro del proyecto TC en el Instituto para lingüística computacional de la Universidad de Stuttgart. Ha sido utilizado con éxito para etiquetar textos en Alemán, Inglés, Francés, Italiano, Español, Griego, y Francés antiguo, y es fácilmente adaptable a otros lenguajes si se dispone de un lexicón y corpus marcado manualmente.

En el área de la minería de opinión también han sido desarrolladas varias herramientas que resuelven alguna etapa de la minería de opinión como la clasificación de la polaridad de las opiniones o la detección de tópicos. Las aplicaciones PosNeg Opinion (Amores, Arco and Artiles, 2015) y OpinionTopicDetection (Orozco, 2016) son una solución a estas etapas de minería de opinión desarrolladas en la UCLV.

PosNeg Opinion

PosNeg Opinion 1.0 es una herramienta que sigue cinco etapas para la detección no supervisada de la polaridad de opiniones en Inglés y Español (Amores, 2013). La Etapa 1 es la encargada de leer las opiniones que fueron especificadas en el XML de entrada y seleccionar los términos que aporten información útil. En la Etapa 2, se parte de cada término que aporta información útil, éste se lematiza y se desambigua lexicalmente. Posteriormente, en la Etapa 3 se traducen los términos seleccionados y se obtienen todas las acepciones del término en inglés. En la Etapa 4 se calcula la polaridad de los términos, considerando la polaridad de cada una de las acepciones del término. Así, en la Etapa 5, al terminar de analizar todos los términos y sus acepciones, la opinión cuenta con un valor positivo y otro negativo, los cuales son comparados, y se toma como polaridad de la opinión el mayor valor (Amores, Arco and Artiles, 2015).

PosNeg Opinion 2.0 tiene el mismo propósito de PosNeg Opinion 1.0, la diferencia radica en que en esta segunda versión se fusionan las etapas 3 y 4 a partir de la aplicación del SpanishSentiWordNet (Amores, Borroto and Arco, 2015) para procesar opiniones en Español sin la necesidad de combinar el SentiWordNet[7] con otras herramientas.

[6] https://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/treetagger.en.html

[7] Recurso léxico para la minería de opinión, asigna puntuaciones de sentimiento a cada synset de WordNet. http://sentiwordnet.isti.cnr.it

Ambas versiones de PosNeg Opinion permiten realizar un análisis de la polaridad de forma local (nivel de oración) y de forma global (nivel de documento).

En el primer caso se realiza por oraciones y se identifican solo las palabras que indican la opinión. Por ejemplo, en su versión 1.0 analiza la siguiente oración "La hp 2000 tiene muy buena batería" y obtiene solo las palabras "muy", "buena" y "batería". Inicialmente se eliminan los términos "la", "hp", "2000" y "tiene" por ser palabras que no ofrecen información para la detección de la polaridad de la opinión. Se buscan las acepciones en Inglés de los términos obtenidos y se suman sus puntuaciones en SentiWordNet para obtener la mayor polaridad, de esta forma se determina que la opinión tiene polaridad positiva.

PosNeg Opinion es capaz de calcular la polaridad de una oración, pero tiene la limitante que no es capaz de determinar cuál es el tópico al que se refiere el criterio emitido, tampoco la entidad y los aspectos que se abordan de la misma.

Supongamos que se desea procesar la siguiente opinión con PosNeg Opinion:

"*El hotel es fantástico y muy elegante. El servicio del conserje es ineficiente. La habitación no estaba limpia cuando llegamos. El restaurante está clasificado como uno de los 3 mejores de la ciudad*."

En este caso PosNeg Opinión calcula la polaridad de toda la opinión mediante el voto global positivo y negativo para la opinión. Las oraciones 1 y 4 tienen polaridad positiva, mientras que las oraciones 2 y 3 tienen polaridad negativa, entonces, ¿cuál será la polaridad de la opinión? ¿Ayudará un valor global de polaridad al gerente del hotel a tomar alguna decisión? Este ejemplo evidencia que PosNeg Opinion no permite calcular la polaridad de las opiniones por tópicos, siendo esta una de sus principales desventajas.

Aunque PosNeg Opinion permite obtener excelentes valores de precisión y exactitud en la detección de la polaridad de las opiniones solo es capaz de clasificar las opiniones expresadas en una oración o la opinión que expresa el texto en su totalidad, y por tanto no realiza un análisis del sentimiento por tópicos. De ahí que se hayan realizado estudios sobre cómo los métodos de detección de tópicos pueden ser útiles en la minería de opinión y contribuir al análisis de sentimientos por tópicos (Arco García, Torres and Amores, 2015).

1.6 Conclusiones finales del capítulo

Existen dos aproximaciones principales para resolver automáticamente la polaridad de un texto: aprendizaje supervisado y aprendizaje no supervisado la cual permite una mejor adaptación a los diferentes dominios y contempla más aspectos del texto.

La minería de opinión constituye un caso particular de la minería de texto de ahí que varias tareas y etapas de la minería de texto deberán ejecutarse para la minería de opinión. Es uno de los temas más complejos del procesamiento de lenguaje natural, a su vez es uno de los más demandados por usuarios e instituciones.

HAC resulta adecuado como método de agrupamiento a emplear en el presente trabajo en su variante de umbral para la construcción de la jerarquía.

La elección de la medida de similitud adecuada constituye un factor primordial en la conformación de grupos.

CAPÍTULO 2. PROCEDIMIENTO GENERAL PARA LA DETECCIÓN DE POLARIDAD POR TÓPICOS DE TEXTOS CORTOS.

En este capítulo se presenta el procedimiento propuesto para la detección de la polaridad por tópicos y se explican cada una de las etapas que conforman este procedimiento. Además, se analizan las particularidades a tener en cuenta para el desarrollo de cada etapa en función del idioma para el cual se implementó el procedimiento, en el caso de este trabajo el idioma seleccionado es el Español.

2.1 Etapas para la detección de la polaridad de opiniones en Español agrupadas por tópicos

Las propuestas que calculan la polaridad a nivel de documento o de oración, brindan una cantidad limitada de información de la opinión, que muchas veces no es suficiente para la toma de decisiones. Por otra parte, el cálculo de la polaridad de las opiniones a nivel de aspecto resulta más complejo, pero permite obtener resultados más útiles en la práctica. El

procedimiento propuesto en este trabajo es una aproximación para el cálculo de la polaridad de las opiniones por tópicos en Español, el esquema general se muestra en la Figura 2.

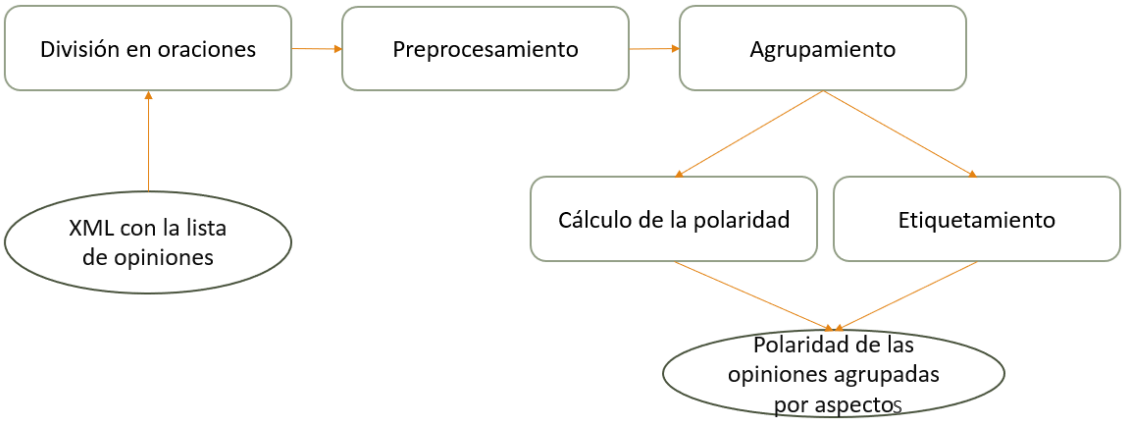

Figura 2: Esquema del procedimiento general propuesto desglosado por etapas

La entrada al procedimiento es una lista de opiniones y la salida los grupos que corresponden con los tópicos que se abordan en las opiniones, etiquetados con la polaridad de cada tópico. Las etapas de procesamiento que se aplican en el procedimiento para producir esta salida se describen a continuación.

2.1.1 Etapa 1: Dividir en oraciones

Las unidades textuales que comúnmente se utilizan para el análisis de textos son las palabras, oraciones, párrafos o los bloques. Estos últimos se refieren a fragmentos en los que se divide el texto; se pueden definir, por ejemplo, especificando una cierta cantidad de tokens o de oraciones. En esta etapa, se seleccionó como unidad textual las oraciones, debido a que los textos de opiniones se caracterizan por presentar un estilo de escritura en forma de composición, con oraciones cortas, escritas de forma informal sin una estructura específica, contenidas en un solo párrafo.

La desambiguación de los límites de las oraciones es una tarea que se estudió con el objetivo de aplicar métodos para identificar los límites de oraciones. Los signos de puntuación son ambiguos, por ejemplo, un punto puede denotar un punto decimal, una abreviación, direcciones de correo electrónico, el fin de una oración, etc. La mayoría de los sistemas usan

gramáticas de expresiones regulares y reglas de excepción para desambiguar los signos de puntuación.

Dos trabajos iniciales fueron los presentados en (Palmer and Hearst, 1994) y (Reynar, Ratnaparkhi and Science, 1997) para trabajar la desambiguación. En el primero se propuso un algoritmo basado en una red neuronal y un vocabulario que contiene información de partes del discurso. En el segundo se presenta una solución basada en un modelo de máxima entropía que no requiere reglas manuales, sino que estima la distribución de probabilidad de un token y el contexto que lo rodea. Para ello utilizan plantillas contextuales como prefijos, sufijos y grados honoríficos, entre otros; luego estiman la probabilidad de identificar los límites.

De forma general, existen tres enfoques para detectar los límites de las oraciones: basados en reglas, basados en técnicas de aprendizaje automático supervisado y no supervisado.

Para esta investigación se seleccionó el detector de oraciones de Apache OpenNLP, que es de tipo supervisado, debido a que utiliza un modelo de entropía máxima para evaluar los caracteres ".", "!", y "?" en una cadena y así determinar si significan el final de la oración o no. Asume que el primer carácter que no está en blanco es el inicio de la primera oración y el último que no está en blanco es el final de la oración. El detector de OpenNLP[8] no requiere *tokenizar* (dividir el texto en palabras o tokens) el texto para detectar las oraciones.

OpenNLP se caracteriza por su fácil integración en una aplicación a través de su API[9] y por poseer buena documentación (Ingersoll, Morton and L.Farris, 2013). En un estudio realizado en (Read *et al.*, 2012), OpenNLP obtuvo el 97.4% de clasificación correcta como promedio en colecciones textuales de distintos dominios (mejor caso 97.6% y peor caso 95.3%). En colecciones de textos informales obtuvo 93.6% (mejor caso 95.1% y peor caso 93.26%).

En forma general, esta etapa consiste en dividir el texto en unidades textuales compactas para su posterior procesamiento. Las opiniones se segmentan por oraciones de punto a punto, para lo cual se hace uso de la biblioteca OpenNLP.

[8] opennlp.tools.sentdetect.SentenceDetectorME
[9] http://opennlp.apache.org/documentation/1.5.3/manual/opennlp.html

2.1.2 Etapa 2: Preprocesar documentos (oraciones)

Para el preprocesamiento del texto se buscaron herramientas que permitieran realizar las técnicas de tokenización y lematización de las palabras, eliminar las palabras que no aportan información significativa del texto y normalizar el texto, es decir, llevar a minúscula las palabras que están en mayúscula. Algunas herramientas de código abierto en el lenguaje de programación Java que implementan estas funciones son Apache Lucene[10], TreeTagger, Stanford CoreNLP y Apache OpenNLP.

Para la tokenización del texto se seleccionó a Apache Lucene porque brinda más flexibilidad que las otras bibliotecas para realizar el análisis léxico de los documentos. Posee diferentes analizadores que permiten eliminar palabras vacías, convertir las palabras a minúsculas y presenta, además, un componente que permite identificar expresiones regulares como direcciones de correo electrónico y fechas. El resto de las herramientas poseen algunas implementaciones para estas funcionalidades, pero no son igual de configurables como Lucene; por ejemplo, Stanford CoreNLP lematiza, pero requiere etiquetar las partes del discurso al mismo tiempo, así como tokenizar y dividir el texto en oraciones, es decir, que las funcionalidades dependen unas de otras.

La herramienta TreeTagger permite realizar el etiquetamiento de partes del discurso y la lematización de forma independiente. En esta etapa solo se aplicaron tanto la lematización como el etiquetado de partes del discurso, de esta forma se elimina la presencia de términos ambiguos, que es necesario para la siguiente etapa de procesamiento. Por tanto, para el preprocesamiento de las opiniones se propone utilizar Apache Lucene y TreeTagger.

El modelo general de preprocesamiento que se aplicó fue la tokenización del texto, convertir todo el texto a minúscula, eliminar las palabras gramaticales (*stop words*), lematizar los términos y agregarle la etiqueta POS (parte del discurso). Ver Figura 3.

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

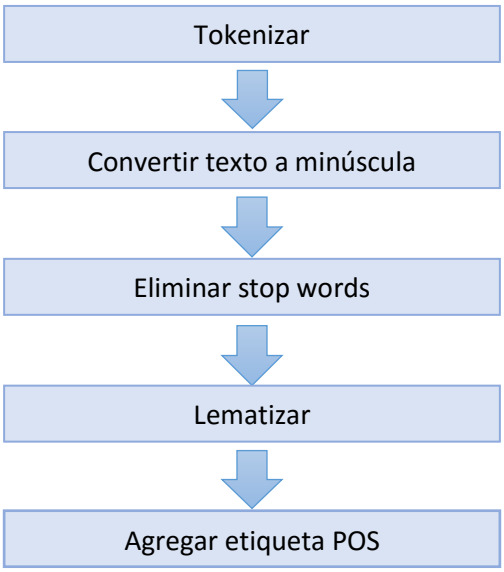

Figura 3: Pasos desarrollados en el preprocesamiento de las oraciones

2.1.3 Etapa 3: Agrupar documentos

El objetivo de los algoritmos de agrupamiento es crear grupos que son coherentes internamente, es decir, que sus objetos o documentos sean lo más similares posibles y a la vez que sean disimilares o difieren de los documentos u objetos de otros grupos (Manning, Prabhakar Raghavan and Schütze, 2008). Para detectar los tópicos es necesario emplear un algoritmo que permita agrupar las oraciones por tópicos, es decir, cada grupo representará un tópico. De esta forma, oraciones que traten un mismo tema serán agrupadas y cada grupo se corresponderá con un tópico de todo el corpus.

En el contexto de la detección de tópicos se han empleado fundamentalmente tres tipos de algoritmos de agrupamiento, los algoritmos jerárquicos (Huang *et al.*, 2013) (Wattenhofer and Cselle, 2007), los particionales (Seo and Sycara, 2004) y los probabilísticos como Expectation-Maximization (Lu *et al.*, 2013).

Los algoritmos de agrupamiento jerárquicos pueden ser aglomerativos (*Hierarchical Agglomerative Clustering*; HAC) o divisivos (Manning, Prabhakar Raghavan and Schütze, 2008). El agrupamiento divisivo ha sido más efectivo para la segmentación de tópicos que para la detección (Purver, 2011).

Dado un conjunto de *N* objetos para agrupar y una matriz de distancia o similitud el algoritmo básico de agrupamiento jerárquico consiste en (Zhu, 2010):

1. Asignar a cada objeto su propio grupo (conformar *N* grupos, cada uno con un solo objeto).
2. Encontrar el par de grupos (más similar) y mezclarlos en un único grupo.
3. Repetir el paso 2 hasta que todos los objetos son agrupados en un único grupo.

Los algoritmos HAC han sido utilizados para detectar tópicos (Huang *et al.*, 2013) (Wattenhofer and Cselle, 2007) y son típicamente visualizados en un dendrograma, Entre ellos se destaca el agrupamiento de enlace único (Single-linkage), agrupamiento de enlace promedio del grupo (Average-linkage) y el agrupamiento de enlace completo (Complete-linkage).

El agrupamiento de enlace único, calcula la similitud entre dos grupos considerando la similitud entre sus dos objetos más similares. Considera un criterio local, es decir, solo tiene en cuenta las áreas donde dos grupos están más cerca uno del otro; las partes más distantes del grupo no se tienen en cuenta (Manning, Prabhakar Raghavan and Schütze, 2008).

El agrupamiento de enlace promedio, calcula la similitud entre dos grupos como el promedio de la similitud entre los pares de objetos de un grupo y otro. Este proceso de enlace promedio es más lento que el agrupamiento de enlace único porque se necesita determinar la similitud promedio entre una gran cantidad de pares de objetos para determinar la similitud del grupo. Por otra parte, es más robusto que enlace único en cuanto a la calidad del agrupamiento (Aggarwal and Zhai, 2012). Se ha recomendado como mejor algoritmo para el agrupamiento de documentos en representaciones vectoriales.

El algoritmo de agrupamiento de enlace completo, calcula la similitud de dos grupos como la similitud de sus miembros más disimilares. No es un criterio local, porque toda la estructura del agrupamiento puede influenciar decisiones de combinación de grupos. Este criterio favorece la obtención de grupos compactos con diámetros pequeños, pero es sensible a objetos que están lejanos. Un solo objeto lejos del centro puede incrementar el diámetro del grupo y cambiar el agrupamiento final. Tiene como desventaja que es sensible a los puntos que no se ajustan en la estructura global del grupo.

Los algoritmos HAC construyen la jerarquía hasta obtener un solo grupo donde se incluyen todos los objetos; sin embargo, en la presente investigación se necesita obtener cierta cantidad de grupos de oraciones que representen los tópicos que se tratan en las opiniones. Por tanto, si se aplican algoritmos HAC es necesario cortar la jerarquía en algún nivel para

obtener una partición. Algunas variantes para obtener una partición a partir del dendrograma son (Manning, Prabhakar Raghavan and Schütze, 2008):

1. Cortar según un nivel predefinido de similitud entre objetos en un mismo grupo.
2. Cortar el dendrograma donde el espacio entre dos combinaciones de similitudes sucesivas entre objetos en un mismo grupo es mayor.
3. Predefinir la cantidad de grupos *K* y seleccionar el punto de corte que produce *K* grupos.
4. Obtener todas las posibles particiones y seleccionar aquella que ofrezca la mejor calidad del agrupamiento, para ello se pueden aplicar medidas de validación internas.

Teniendo en cuenta que no resulta trivial hacer un corte en el dendograma para obtener los grupos, se pudiera pensar en la aplicación de un algoritmo de agrupamiento plano que obtenga directamente una partición. El algoritmo K-means (Liaoning, 2013), por ejemplo, es un algoritmo plano partitivo clásico y es más eficiente que los algoritmos jerárquicos; sin embargo, tiene varias desventajas, entre ellas: crea grupos sin una estructura explícita que los relacione, es necesario especificar la cantidad de grupos a obtener y es sensible al conjunto inicial de semillas escogidas durante el agrupamiento.

En esta investigación no se cuenta con conocimiento a priori que permita especificar los parámetros que requieren la mayoría de los algoritmos planos partitivos en su inicialización, de ahí que se sugiere utilizar los algoritmos HAC.

La aplicación de los algoritmos HAC impone definir cuál variante utilizar para obtener una partición a partir de la jerarquía. Para lograr un punto de corte estándar para una variabilidad en la longitud de las opiniones, se aplica un umbral que permita agrupar comparando las medidas de similitud de los grupos con dicho umbral. Los grupos se conformarán hasta que la mayor similitud de un grupo sea menor que el umbral especificado, si es igual o mayor se detiene el agrupamiento.

Para realizar este enfoque se propone el uso de cuatro expresiones para calcular los umbrales (Shulcloper, 2010): la media de las similitudes entre todos los pares de objetos posibles, la media de los valores máximos de las similitudes entre cualquier par de objetos, la media de los valores mínimos de las similitudes entre cualquier par de objetos y la media ponderada de la media de las similitudes y la media de los máximos (Arco, 2008).

Para calcular la similitud entre documentos se propone la medida de similitud propuesta en (Mihalcea, Corley and Strapparava, 2006), que se expresa en la ecuación (2). Donde T_1 y T_2 son las dos oraciones que se están comparando y *w* es un término que pertenece a T_1 o T_2, $maxSim(w, T_i)$ es calculado como la similitud entre *w* y la palabra más cercana a ella en la otra oración según la medida de similitud entre palabras propuesta en (Jiang and Conrath, 1997).

$$\text{sim}(T_1, T_2) = \frac{1}{2}\left(\frac{\sum_{w\epsilon\{T_1\}}\left(maxSim(w, T_2) * idf(w)\right)}{\sum_{w\epsilon\{T_1\}} idf(w)} + \frac{\sum_{w\epsilon\{T_2\}}\left(maxSim(w, T_1) * idf(w)\right)}{\sum_{w\epsilon\{T_2\}} idf(w)}\right) \quad (2)$$

Esta medida, al tener en cuenta información semántica es más adecuada para textos cortos como lo son las opiniones, en la etapa de evaluación se comprobará que brinda mejores resultados que las medidas de similitud que tienen en cuenta solamente la información sintáctica.

2.1.4 Etapa 4: Etiquetar grupos

El etiquetado de tópicos (topic labeling) se refiere a otorgarle a cada grupo de oraciones una breve descripción que identifique el tópico que tratan. Una etiqueta puede estar constituida por una o varias palabras que mejor representen un tópico.

Existen métodos no supervisados que aplican una distancia para determinar los términos con mayores puntuaciones y estos conforman la etiqueta del grupo (Xu and Oard, 2011).

Algunos enfoques de etiquetamiento de tópicos ordenan las palabras más relevantes de un segmento de tópico y con ellas construyen las etiquetas; por ejemplo, se enfocan en obtener descripciones de las relaciones jerárquicas entre los tópicos, apoyándose en la extracción de frases significativas utilizando técnicas de análisis de secuencias de textos y generación de n-grams (Mao, 2012; Mei, Shen and Zhai, 2007).

El método de LDA puede usarse también para la tarea de etiquetamiento de tópicos (Carenini, Murray and Ng, 2011; Hingmire, 2013). Una propuesta interesante es un método no supervisado basado en un grafo y que emplea el algoritmo PageRank para pesar las

palabras (Aletras and Stevenson, 2014). La puntuación final de las etiquetas candidatas es la suma de los pesos de sus palabras. La etiqueta con la mayor puntuación se selecciona para representar el tópico.

Seleccionar etiquetas o tópicos es más difícil que escoger objetos representativos de un grupo. La mayor parte de los enfoques se dirigen a encontrar términos importantes y frases en el grupo. Uno de los métodos propuestos por investigadores (Tagarelli and Karypis, 2012), consiste en otorgar pesos a los términos; por ejemplo, a partir de TF-IDF y devolver una lista de términos ordenados por pesos.

No todos los métodos de etiquetamiento de grupos obtienen etiquetas que constituyen los tópicos que aborda el grupo; por ejemplo, aquellos métodos que identifican el documento más cercano al centro del grupo efectivamente lo caracterizan, pero no representa un tópico. Esto sucede con varios algoritmos de etiquetamiento; de ahí que detectar tópicos en grupos de oraciones impone retos mayores al etiquetamiento de grupos.

Teniendo en cuenta que los sustantivos son los que aportan más información acerca del tópico que se trata en el grupo, se tuvieron en cuenta solamente este tipo de palabras para conformar la etiqueta.

P1: Extraer los sustantivos del grupo, los cuales están previamente anotados desde la etapa de preprocesamiento.

P2: Para cada uno de los sustantivos calcular la medida tf-idf

P3: Ordenar los sustantivos por su peso.

P4 Tomar los primeros sustantivos ordenados.

Los sustantivos que conforman la etiqueta corresponden con los tópicos que se tratan en el grupo.

2.1.5 Etapa 5: Calcular polaridad

Una vez conformados los grupos se pasa a calcular la polaridad de cada una de las oraciones correspondientes a los grupos y a su combinación para determinar la polaridad total de los grupos.

La polaridad de las oraciones se obtiene a partir de la polaridad de cada una de las palabras que la conforman. Los valores de polaridad negativa y positiva de los términos se calculan a partir del recurso SpanishSentiWordNet (Amores, Borroto and Arco, 2015).

En SpanishSentiWordNet aparecen las puntuaciones negativas y positivas de cada una de las acepciones de los términos. En este trabajo, solo se tiene en cuenta las que corresponden con la parte del discurso con la que fue anotada la palabra en la etapa de preprocesamiento, de esta forma se disminuye la posibilidad de obtener una puntuación incorrecta debido a la ambigüedad léxica.

La polaridad positiva y negativa de las oraciones se calculan como se expresan en las ecuaciones ((*3*) y ((*4*) respectivamente, siendo T_i cada uno de los términos de la oración

$$PosOracion_j = \sum_{T_i \in Oracion_j} Pos(T_i) \quad (3)$$

$$NegOracion_j = \sum_{T_i \in Oracion_j} Neg(T_i) \quad (4)$$

Mientras que la polaridad del grupo vendrá expresada por la suma de las polaridades positiva o negativa de las oraciones que conforman el grupo. Las ecuaciones ($PosGrupo_k = Oracion_j \in Grupo_k$ $PosOracion_j$ *(5)* y $NegGrupo_k = \sum_{Oracion_j \in Grupo_k} NegOracion_j$ (6) muestran cómo se obtienen estas.

$PosGrupo_k = \sum_{Oracion_j \in Grupo_k} PosOracion_j$ (5)

$NegGrupo_k = \sum_{Oracion_j \in Grupo_k} NegOracion_j$ (6)

2.2 Conclusiones parciales del capítulo

Se presenta un procedimiento general para calcular la polaridad de opiniones por tópicos en Español. Para la detección de los tópicos se propone realizar un agrupamiento jerárquico

aglomerativo calculando la similitud entre documentos con una medida de similitud semántica, esto permite que se obtenga un agrupamiento de mayor calidad, por lo tanto, la detección de tópicos es más acertada. El etiquetado de los grupos se realiza con la determinación de los sustantivos más representativos, y estos identifican los tópicos tratados en el grupo.

El procedimiento es aplicable para cualquier idioma, solo es necesario sustituir los recursos utilizados por los específicos del idioma para el cual se implemente. Esto hace que el procedimiento sea además de independiente del dominio, muy sencillo de implementar para varios idiomas.

CAPÍTULO 3. POSNEGTOPICOPINION. IMPLEMENTACIÓN COMPUTACIONAL DE LA DETECCIÓN DE POLARIDAD DE TEXTOS EN ESPAÑOL ORIENTADA A TÓPICOS

En este capítulo se presenta la biblioteca PosNegTopicOpinion que se desarrolló siguiendo el procedimiento propuesto para el cálculo de la polaridad de las opiniones en español orientada a tópicos. Se detallan los elementos de diseño y requerimientos generales del software. Finalmente se presentarán los resultados obtenidos en la evaluación de la herramienta que validan la eficacia del procedimiento propuesto.

3.1 Diagramas fundamentales

A partir del procedimiento propuesto se desarrolló la biblioteca PosNegTopicDetection para el cálculo de la polaridad de opiniones en español orientada a tópicos. Esta biblioteca consta

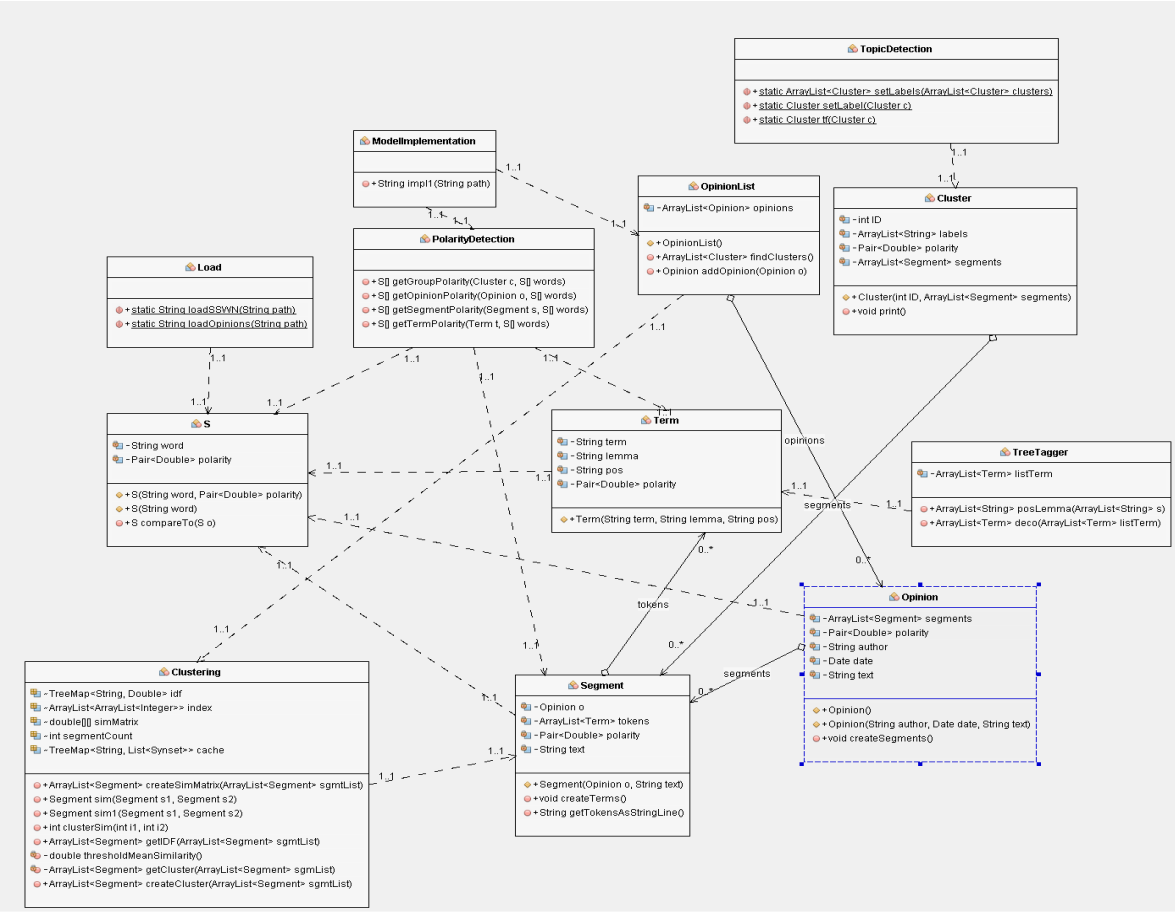

Figura 4: Diagrama de clases de PosNegTopicDetection

de 12 clases implementadas en un solo paquete de la forma que se muestra en la Figura 4: Diagrama de clases de PosNegTopicDetection.

La clase `ModelImplementation` es una clase controladora, contiene el método impl1 donde se ejecutan cada uno de los pasos del procedimiento.

La clase `Load` contiene el método `loadOpinions` como se muestra en la Figura 5, que carga el fichero XML, el cual contiene las opiniones a procesar. El método `loadSSWN` carga el recurso SpanishSentiWordNet para el cálculo de la polaridad.

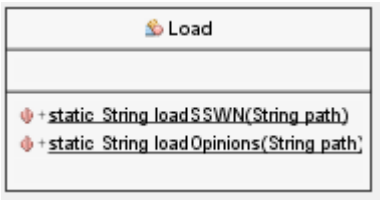

Figura 5: Clase Load

La clase `Segment` cuenta con el método `createTerm` como se muestra en la Figura 6, que realiza el preprocesamiento de la oración y guarda en el atributo tokens la lista de términos ya lematizados y anotados con la etiqueta POS. Para esta tarea hace uso de la clase `TreeTagger`.

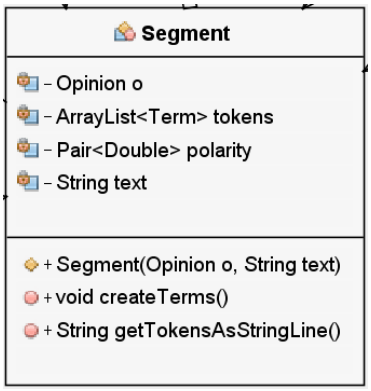

Figura 6: Clase Segment

La clase `OpinionList` se instancia con la lista de opiniones extraídas del fichero XML. Esta clase contiene el método `findCluster` como se muestra en la Figura 7 donde se instancia un objeto de la clase `Clustering` que contiene los métodos para la conformación

de grupos con la implementación de un agrupamiento jerárquico aglomerativo siguiendo la estrategia average linkage como se muestra en la Figura 8.

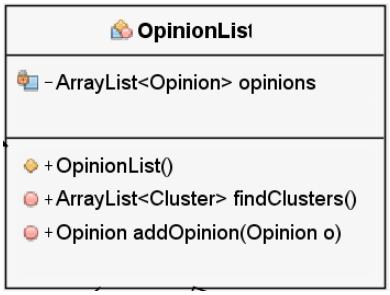

Figura 7: Clase OpinionList

La clase `PolarityDetection` contiene los métodos para el cálculo de la polaridad de las opiniones a los distintos niveles (grupo, opinión, oración y término) como se muestra en la Figura 9, esta forma de implementación permite obtener los resultados de la polaridad de forma independiente.

La clase `TopicDetection` contiene el método `setLabel` que se encarga de etiquetar el grupo (parámetro c) con los sustantivos contenidos en el mismo a partir del valor de tf (term frecuency) que se calcula en el método tf como se muestra en la Figura 10.

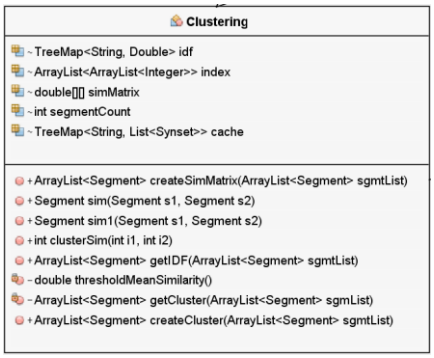

Figura 8: Clase Clustering

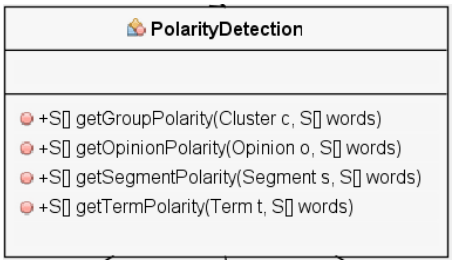

Figura 9: Clase PolarityDetection

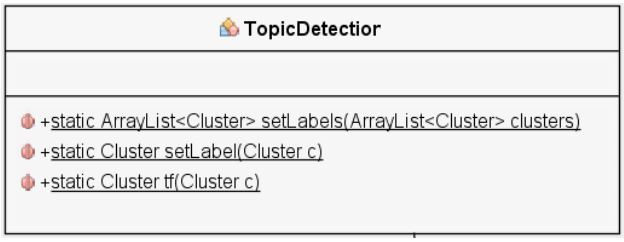

Figura 10: Clase TopicDetection

Para la implementación de la biblioteca se utilizaron varias librerías y recursos externos. En el diagrama de componentes que se muestra en la Figura 11 se puede observar la composición de la biblioteca con respecto a sus dependencias.

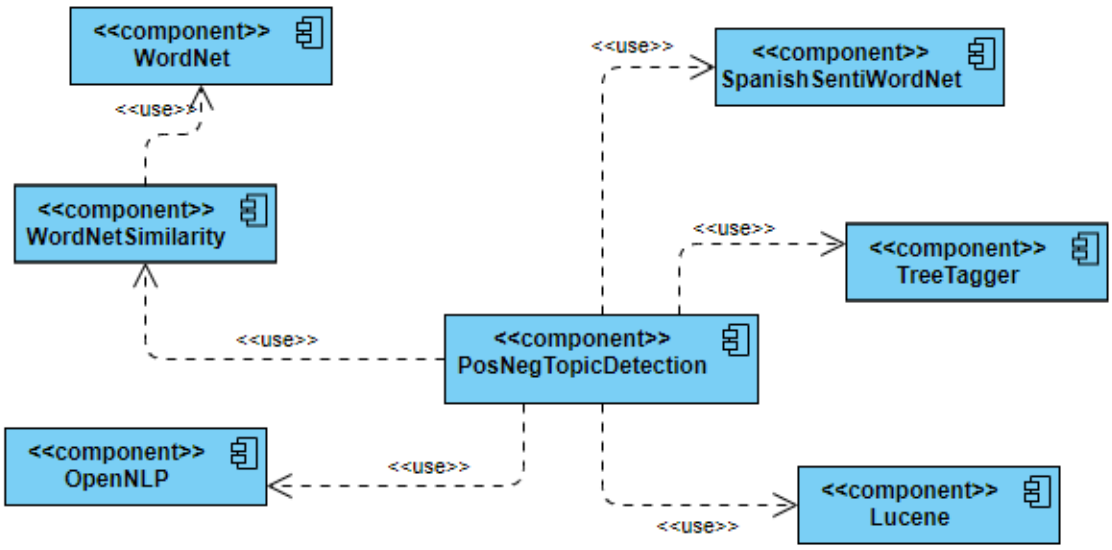

Figura 11: Diagrama de componentes

La biblioteca PosNegTopicOpinion se implementó en el lenguaje Java lo que permite utilizarla en cualquier sistema operativo.

TreeTagger es el único componente que debe modificarse en dependencia del sistema operativo en que se use la biblioteca, pues existen distribuciones diferentes para Windows y Linux.

3.2 Limitaciones de la implementación

La principal limitación de la implementación de la biblioteca radica en el tiempo de ejecución, debido a la alta complejidad computacional que tiene el cálculo de la similitud semántica entre documentos, por lo que la etapa de agrupamiento se hace más lenta.

Otra limitación que puede afectar la eficacia de la biblioteca es que en la versión actual no se tiene en cuenta el efecto de la negación en la polaridad de las opiniones.

3.3 Evaluación de la herramienta

Para la validación del esquema propuesto se utilizó el corpus COAH. El contenido de este corpus son opiniones de usuarios reales de hoteles andaluces, cada opinión tiene una puntuación de 1 a 5 que representa la calidad del hotel según el usuario, el identificador de la opinión y el texto de la opinión. En la Tabla 2 se muestran las características principales del corpus.

Formato	XML
Número de opiniones	1816
Número de tokens	272446
Longitud media de las oraciones	23.245

Tabla 2: Características del corpus

El corpus se dividió en 6 fracciones más pequeñas, la selección de las opiniones en cada una de las fracciones se realizó de manera aleatoria. Esta división se hizo con el objetivo de disminuir el tiempo de ejecución de la biblioteca y además tener más variedad en la muestra para la evaluación.

La estrategia trazada fue la evaluación de cada uno de los módulos del procedimiento de forma independiente y la unión de estos resultados se tomó como índice de calidad del procedimiento propuesto.

La etapa de división en oraciones se realizó con la librería OpenNLP que brinda para el idioma español un **97.52%** de eficacia en esta tarea. Para el corpus utilizado se comprobó la calidad de las divisiones con el objetivo de conocer en qué medida se ve afectado el funcionamiento de la librería OpenNLP con textos que contienen errores. Se comprueba que no se afecta el rendimiento.

El etiquetado de los grupos fue evaluado teniendo en cuenta la lógica de las etiquetas asignadas en función de los tópicos reales de las oraciones pertenecientes al grupo. Para las pruebas realizadas las etiquetas estuvieron en concordancia con los principales temas tratados en los grupos, aunque en ocasiones son repetitivas ya que aparecen varias con el mismo significado.

Para la evaluación de la polaridad se calculó el porciento de clasificaciones correctas de acuerdo a las etiquetas de las opiniones. El corpus utilizado para la evaluación tiene cada una de las opiniones etiquetadas de uno a cinco, de manera que mientras más alto es el valor de la puntuación, más positiva es, se asumió que de uno a tres la polaridad es negativa, y cuatro y cinco corresponden con una opinión positiva. Por otro lado, se calculó la polaridad positiva y negativa de cada opinión con la biblioteca y se clasificó de acuerdo a la mayor polaridad. De esta forma se obtuvo el porciento de exactitud promedio de la clasificación de 74%. Este resultado se ve afectado principalmente por el lenguaje informal que caracteriza al corpus utilizado para la evaluación lo que hace que existan muchos términos que no se hallen en el SpanishSentiWordNet.

La etapa de agrupamiento fue la etapa que se le prestó mayor interés en la evaluación debido a que es la de más influencia en la calidad del procedimiento. Además, es la etapa más compleja computacionalmente debido a la naturaleza corta de las opiniones. Cuando se tratan textos cortos se tiene menor cantidad de información, lo que hace más difícil su caracterización y por tanto tiende a verse afectada la calidad del agrupamiento.

Para realizar la evaluación del agrupamiento se aplicó el índice de Silhouette como medida de calidad.

Inicialmente la medida de similitud aplicada para el agrupamiento fue la medida de similitud Coseno y se probaron dos variantes para el cálculo del umbral: la media de las similitudes y la media de los valores máximos de las similitudes, en la Tabla 3 se muestran los resultados de esta evaluación.

Corpus	**Umbral**	**Valor de Silhouette**
Data_01	Media de las similitudes	-0,04
Data_01	Media de los valores máximos de las similitudes	-0,08
Data_02	Media de las similitudes	-0,09
Data_02	Media de los valores máximos de las similitudes	-0,12

Tabla 3: Evaluación del agrupamiento utilizando similitud Coseno

Como se puede observar los resultados de la evaluación no son los más deseables en cuanto al índice de Silhuette, estos resultados se deben mayormente a que los textos son muy cortos por lo que no brindan suficiente información sintáctica para obtener un agrupamiento coherente.

Como resultado de la evaluación anterior, se concluye que no es posible aplicar un agrupamiento que utilice solamente la información sintáctica, por tanto, se sustituyó la medida de similitud por la propuesta por (Mihalcea, Corley and Strapparava, 2006). Los resultados de esta nueva evaluación se muestran en la Tabla 4.

Corpus	**Umbral**	**Valor de Silhouette**
Data_01	Media de las similitudes	0,31
Data_01	Media de los valores máximos de las similitudes	0,62
Data_02	Media de las similitudes	0,38
Data_02	Media de los valores máximos de las similitudes	0,64

Tabla 4: Evaluación del agrupamiento utilizando similitud semántica

Los resultados de la evaluación utilizando similitud semántica son superiores con respecto a los que se obtuvieron con la medida de similitud Coseno. Se puede apreciar además en esta tabla que el umbral que se calcula teniendo en cuenta la media de los valores máximos de similitud ofrece mejores resultados. Esto se debe principalmente a que la generación de grupos se detiene más rápido, quedando de esta manera, grupos más pequeños y compactos, lo que hace que el efecto de atracción que ejercen los grupos más grandes hacia los más pequeños tenga menor impacto en la calidad del agrupamiento.

3.4 Conclusiones parciales

Se implementó la biblioteca PosNegTopicDetection que implementa el procedimiento propuesto para la detección de polaridad orientada a tópicos de textos cortos en Español. La biblioteca se implementó en JAVA , es portable, permite realizar el minado de las opiniones no dependientes de un dominio específico, lo cual facilita su aplicabilidad.

La evaluación de la implementación verifica la factibilidad de la misma y sugiere la necesidad de buscar eficiencia en la etapa de conformar los grupos de opiniones.

CONCLUSIONES

Se identificaron e integraron herramientas que facilitaron la implementación de las etapas del procedimiento propuesto. Cada una de las herramientas y recursos fueron seleccionados en base a la calidad de los resultados que ofrecen y a la facilidad de integración entre sí.

Se diseña un procedimiento para el cálculo de la polaridad por tópicos de textos cortos, independiente del dominio y aplicable a varios idiomas que consta con cinco etapas, cada una de ellas constituye un módulo independiente lo que permite que pueda modificarse cada etapa sin afectar las restantes. El etiquetado y el cálculo de la polaridad se pueden ejecutar en paralelo.

Se implementó una biblioteca que integra los diferentes módulos del procedimiento y recursos que facilitan el análisis para textos en español, es portable, permite realizar el minado de las opiniones no dependientes de un dominio específico, lo cual facilita su aplicabilidad.

El procedimiento propuesto fue evaluado aplicando distintos criterios de prueba a cada una de las etapas implementadas en la biblioteca. Para el agrupamiento se obtuvo un valor máximo de índice de Silhouette de 0.64, lo que demuestra la coherencia de los grupos obtenidos. Se alcanzó un promedio de exactitud de 74% en la clasificación de la polaridad, dado principalmente por la escritura informal de las opiniones. El etiquetado de los grupos en la mayoría de los casos estuvo acorde a los principales temas tratados en el grupo.

RECOMENDACIONES

Paralelizar el módulo de agrupamiento para mejorar la eficiencia computacional que se ve afectado por el costo del cálculo de la similitud semántica.

Incluir el análisis de las palabras negadoras en la opinión para aumentar la exactitud en la etapa de cálculo de la polaridad.

REFERENCIAS BIBLIOGRÁFICAS

Aggarwal, C. C. and Zhai, C. (2012) *Mining Text Data*. Springer.

Aletras, N. and Stevenson, M. (2014) 'Labelling Topics using Unsupervised Graph-based methods', in *ACL Short Papers*, p. 661.

Allan, J. (2002) *Topic Detection and Tracking: Event-Based Information Organization*. doi: 10.1007/978-1-4615-0933-2.

Amores, M. (2013) *Detección no supervisada de la polaridad de las opiniones*. Universidad Central 'Marta Abreu' de Las Villas.

Amores, M., Arco, L. and Artiles, M. (2015) 'PosNeg opinion : Una herramienta para gestionar comentarios de la Web', *Revista Cubana de Ciencias Informáticas*, 9(1), pp. 20–12.

Amores, M., Borroto, C. and Arco, L. (2015) 'SentiWordNet 4.0 and SpanishSenti-WordNet assisting Polarity Detection', *Eureka Workshop*.

Arco García, L., Torres, C. and Amores, M. (2015) 'Propuesta de incorporación de técnicas de detección de tópicos a PosNeg Opinion', (October).

Arco, L. (2008) *Agrupamiento basado en la intermediación diferencial y su valoración utilizando la teoría de los conjuntos aproximados*. Universidad Central 'Marta Abreu' de Las Villas.

Baccianella, S., Esuli, A. and Sebastiani, F. (2010) 'SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining.', in *LREC*, pp. 2200–2204.

Bahrainian, S.-A. and Dengel, A. (2013) 'Sentiment analysis and summarization of twitter data', in *Computational Science and Engineering (CSE), 2013 IEEE 16th International Conference on*, pp. 227–234.

Beineke, P. *et al.* (2003) 'An exploration of sentiment summarization', in *Proceedings of AAAI*, pp. 12–15.

Berry, M. W. (2004) *Survey of Text Mining*.

Blair-Goldensohn, S. *et al.* (2008) 'Building a sentiment summarizer for local service reviews', in *WWW Workshop on NLP in the Information Explosion Era*, pp. 339–348.

Carenini, G., Cheung, J. C. K. and Pauls, A. (2013) 'MULTI-DOCUMENT SUMMARIZATION OF EVALUATIVE TEXT', *Computational Intelligence*. Wiley Online Library, 29(4), pp. 545–576.

Carenini, G., Murray, G. and Ng, R. (2011) *Methods for Mining and Summarizing Text Conversations*.

Chen, Z. and Liu, B. (2016) 'Lifelong Machine Learning', *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 10(3), pp. 1–145. doi: 10.2200/S00737ED1V01Y201610AIM033.

Das, D. and Martins, A. F. T. (2007) 'A survey on automatic text summarization', *Literature Survey for the Language and Statistics II course at CMU*, 4, pp. 192–195.

Demey, J. R. *et al.* (2015) 'Medidas de distancia y de similitud', *Valoración y análisis de la diversidad funcional y su relación con los servicios ecosistémicos*, (December 2015), pp. 47–59.

Dong, H., Hui, S. C. and He, Y. (2006) 'Structural Analysis of Chat Messages for Topic Detection', *Online Information Review*.

Esuli, A. (2008) 'Automatic Generation of Lexical Resources for Opinion Mining : Models , Algorithms and Applications'.

Esuli, A. and Sebastiani, F. (2005) 'Determining the semantic orientation of terms through gloss classification', in *Proceedings of the 14th ACM international conference on Information and knowledge management*, pp. 617–624.

Esuli, A. and Sebastiani, F. (2006a) 'Determining Term Subjectivity and Term Orientation for Opinion Mining.', in *EACL*, p. 2006.

Esuli, A. and Sebastiani, F. (2006b) 'Sentiwordnet: A publicly available lexical resource for opinion mining', in *Proceedings of LREC*, pp. 417–422.

Esuli, A. and Sebastiani, F. (2007) *SENTIWORDNET: A high-coverage lexical resource for*

opinion mining.

Esuli, A. and Sebastiani, F. (2009) 'Enhancing opinion extraction by automatically annotated lexical resources', in *Human Language Technology. Challenges for Computer Science and Linguistics*. Springer, pp. 500–511.

Fellbaum, C. (1998) *WordNet*. Wiley Online Library.

Ganesan, K., Zhai, C. and Han, J. (2010) 'Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions', in *Proceedings of the 23rd international conference on computational linguistics*, pp. 340–348.

Hingmire, S. (2013) 'Document Classification by Topic Labeling', *SIGIR*, pp. 877–880.

Hu, M. *et al.* (2010) 'Opinion Extraction, Summarization and Tracking in News and Blog Corpora.', in *AAAI spring symposium: Computational approaches to analyzing weblogs*. Citeseer, pp. 261–377.

Huang, X. *et al.* (2013) 'A Topic Detection Approach Through Hierarchical Clustering on Concept Graph', *Applied Mathematics & Information Sciences*, 2295(6), pp. 2285–2295.

Hutchison, D. (2013) *Computational Linguistics and Intelligent Text Processing*.

Ingersoll, G. S., Morton, T. S. and L.Farris, A. (2013) *Taming Text*.

Jiang, J. and Conrath, D. (1997) 'Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy'. doi: 10.1152/ajplegacy.1959.196.2.457.

Kaur, A. and Duhan, N. (2015) 'A Survey on Sentiment Analysis and Opinion Mining', *International Journal of Innovations & Advancement in Computer Science*, 4(May), pp. 107–116. Available at: https://scholar.google.com.br/scholar?start=140&q=tudonotítulo:+(Emotion+OR+Feeling+OR+Sentiment+OR+Opinion+OR+Personality+OR+Subjectivity)+AND+Mining&hl=pt-BR&as_sdt=0,5&as_ylo=2005&as_yhi=2015#12.

Khurana, D. *et al.* (2017) 'Natural Language Processing : State of The Art , Current Trends and Challenges', (August 2017).

Kim, H. D. and Zhai, C. (2009) 'Generating comparative summaries of contradictory

opinions in text', in *Proceedings of the 18th ACM conference on Information and knowledge management*, pp. 385–394.

Kim, S. *et al.* (2004) 'Determining the Sentiment of Opinions'.

KumarPatidar, A., Agrawal, J. and Mishra, N. (2012) 'Analysis of Different Similarity Measure Functions and Their Impacts on Shared Nearest Neighbor Clustering Approach', *International Journal of Computer Applications*, 40, pp. 1–5. doi: 10.5120/5061-7221.

Lerman, K., Blair-Goldensohn, S. and McDonald, R. (2009) 'Sentiment summarization: evaluating and learning user preferences', in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 514–522.

Lerman, K. and McDonald, R. (2009) 'Contrastive summarization: an experiment with consumer reviews', in *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics, companion volume: Short papers*, pp. 113–116.

Liaoning, T. (2013) 'The K-Means Clustering Algorithm', 48(2), pp. 762–767.

Liu, B. (2010) 'Sentiment Analysis and Subjectivity', pp. 1–38.

Liu, B. (2012) *Sentiment Analysis and Opinion Mining*. doi: 10.1007/978-1-4899-7502-7_907-1.

Liu, B. (2015) *Sentiment Analysis*. doi: 10.1109/ICETACS.2013.6691379.

Liu, B. (2017) 'Opinion Mining', *Encyclopedia of Database Systems*, (1), pp. 1–6. doi: 10.1007/978-1-4899-7993-3_257-2.

Liu, C.-L. *et al.* (2012) 'Movie rating and review summarization in mobile environment', *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*. IEEE, 42(3), pp. 397–407.

Lu, Y. *et al.* (2010) 'Exploiting structured ontology to organize scattered online opinions', in *Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 734–742.

Lu, Y. *et al.* (2013) 'Health-Related Hot Topic Detection in Online Communities Using

Text Clustering', *PLoS ONE*, 8(2), pp. 1–9. doi: 10.1371/journal.pone.0056221.

Magdaleno Guevara, D. (2015) *Metodología para el agrupamiento de documentos semiestructurados*.

Manning, C., Prabhakar Raghavan and Schütze, H. (2008) *An Introduction to Information Retrieval*. Cambridge University Press.

Mao, X. (2012) 'Automatic Labeling Hierarchical Topics', *CIKM*.

Mihalcea, R., Corley, C. and Strapparava, C. (2006) 'Corpus-based and Knowledge-based Measures of Text Semantic Similarity'.

Miller, G. A. *et al.* (1990) 'Introduction to wordnet: An on-line lexical database*', *International journal of lexicography*. Oxford Univ Press, 3(4), pp. 235–244.

Miller, G. A. (1995) 'WordNet: a lexical database for English', *Communications of the ACM*. ACM, 38(11), pp. 39–41.

Moens, M.-F., Li, J. and Chua, T.-S. (2014) *Mining User Generated Content, Mining User Generated Content*.

Nishikawa, H. *et al.* (2010a) 'Opinion summarization with integer linear programming formulation for sentence extraction and ordering', in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pp. 910–918.

Nishikawa, H. *et al.* (2010b) 'Optimizing informativeness and readability for sentiment summarization', in *Proceedings of the ACL 2010 Conference Short Papers*, pp. 325–330.

Orozco, C. (2016) *Segmentación y detección de tópicos enfocado a la minería de opinión*.

Palmer, D. D. and Hearst, M. A. (1994) 'Adaptive Sentence Boundary Disambiguation', *Proceedings of the 4th Conference on Applied Natural Language Processing*, (2).

Park, S., Lee, K. and Song, J. (2011) 'Contrasting opposing views of news articles on contentious issues', in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pp. 340–349.

Pascual, D., Pla, F. and Sánchez, S. (2007) 'Algoritmos de agrupamiento', *Métodos*

informáticos avanzados, pp. 163–175. Available at: http://marmota.dlsi.uji.es/WebBIB/papers/2007/1_Pascual-MIA-2007.pdf.

Paul, M. J., Zhai, C. and Girju, R. (2010) 'Summarizing contrastive viewpoints in opinionated text', in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 66–76.

Petkos, G., Aiello, L. and Skraba, R. (2014) 'A soft frequent pattern mining approach for textual topic detection', *WIMS*.

Purver, M. (2011) 'Topic Segmentation', in *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, pp. 1–28.

Ranade, S. *et al.* (2013) 'Online debate summarization using topic directed sentiment analysis', in *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, p. 7.

Raut, V. B. and Londhe, D. D. (2014) 'Survey on Opinion Mining and Summarization of User Reviews on Web', *IJCSIT) International Journal of Computer Science and Information Technologies*. Citeseer, 5(2), pp. 1026–1030.

Read, J. *et al.* (2012) 'Sentence Boundary Detection : A Long Solved Problem?', *COLING*, (December 2012), pp. 985–994.

Reynar, J. C., Ratnaparkhi, A. and Science, I. (1997) 'A Maximum Entropy Approach to Identifying Sentence Boundaries', *Proceedings of the fifth conference on Applied natural language processing. Association for Computational Linguistics*, pp. 16–19.

Seki, Y. *et al.* (2006) 'Opinion-focused summarization and its analysis at DUC 2006', in *Proceedings of the Document Understanding Conference (DUC)*, pp. 122–130.

Seo, Y. and Sycara, K. (2004) *Text clustering for topic detection*. Pittsburgh, Pennsylvania.

Sharma, N. R. and Chitre, V. D. (2014) '2014 Opinion Mining, analysis and its challenges.pdf', *International Journal of Innovations & Advancement in Computer Science*, p. 7.

Shulcloper, J. R. (2010) *Reconocimiento lógico combinatorio de patrones: teoría y*

aplicaciones.

Steinberger, J. (2013) *Multilingual Summarisation and Sentiment Analysis*. Citeseer.

Tagarelli, A. and Karypis, G. (2012) 'A segment-based approach to clustering multi-topic documents', *Knowledge and information systems*, 34(3), pp. 563–595. doi: 10.1007/s10115-012-0556-z.

Tata, S. and Di Eugenio, B. (2010) 'Generating fine-grained reviews of songs from album reviews', in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 1376–1385.

Titov, I. and Mcdonald, R. (2008) 'Modeling Online Reviews with Multi-grain Topic Models', *Proceedings of the 17th international conference on World Wide Web. ACM.*, pp. 111–120.

Vasantharaj, S. *et al.* (2015) 'A Survey on Sentiment Analysis Applied in Opinion Mining', *Journal of Network Communications and Emerging Technologies*, 1(1), pp. 15–21. doi: 10.1016/j.joms.2009.03.049.

Wang, D. and Liu, Y. (2011) 'A pilot study of opinion summarization in conversations', in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pp. 331–339.

Wattenhofer, R. and Cselle, G. (2007) 'BuzzTrack : Topic Detection and Tracking in Email', *IUI*.

Xu, T. and Oard, D. W. (2011) 'Wikipedia-based Topic Clustering for Microblogs', *ASIST*.

Yatani, K. *et al.* (2011) 'Analysis of adjective-noun word pair extraction methods for online review summarization', in *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, p. 2771.

Ye, J. *et al.* (2006) *Protej: Biomedical Topic Detection and Tracking*.

Zhang, L. and Liu, B. (2014) 'Aspect and Entity Extraction for Opinion Mining', in *Data Mining and Knowledge Discovery for Big Data*. Springer Berlin Heidelberg, pp. 1–40.

Zhang, Z. (2012) 'K-means Algorithm Cluster Analysis in Data Mining', p. 3. Available at:

http://user.engineering.uiowa.edu/~ie_155/lecture/K-means.pdf.

Zhu, X. (2010) ‘CS769 Spring 2010 Advanced Natural Language Processing - Clustering’, pp. 1–4.

Zhuang, L., Jing, F. and Zhu, X.-Y. (2006) ‘Movie review mining and summarization’, in *Proceedings of the 15th ACM international conference on Information and knowledge management*, pp. 43–50.