UNIVERSITAT DE BARCELONA

Análisis del Género Discursivo Aplicado a la Clasificación Automática de la Polaridad en Comentarios sobre Productos

John Alexander Roberto Rodríguez

Análisis del Género Discursivo Aplicado a la Clasificación Automática de la Polaridad en Comentarios sobre Productos

John Alexander Roberto Rodríguez

Tesis presentada para optar al grado de Doctor en Lingüística
en el programa de doctorado en *Lingüística teórica, computacional y aplicada*,
Departamento de Lingüística,
Universidad de Barcelona

bajo la supervisión de

Dra. Maria Salamó Llorente
Departamento de Matemática Aplicada y Análisis
Universidad de Barcelona

Dra. Maria Antònia Martí Antonín
Departamento de Lingüística
Universidad de Barcelona

Tutor: Dra. Maria Antònia Martí Antonín

Octubre de 2015

AGRADECIMIENTOS

Quiero expresar mi agradecimiento a las personas que han contribuido directa o indirectamente en el desarrollo de esta tesis.

En primer lugar a mis directoras de tesis, Maria Salamó y Maria Antònia Martí, por su dedicación y experiencia. A los miembros del *Centre de Llenguatge i Computació* (CLiC) y del *Servei de Tecnologia Lingüística* (STeL) por su colaboración y apoyo: Mariona Taulé Delor, Montserrat Nofre, Aina Peris Morant, Marta Vila Rigat, Glòria de Valdivia Pujol, Marta Recasens, Santi Reig y Oriol Borrega.

En segundo lugar, también quiero agradecer a las personas que desde fuera del ámbito académico me han brindado su apoyo constante: a mis padres, a Maribel y, en general, a mis familias, colombiana y catalana.

Esta tesis ha sido financiada por una beca de la *Generalitat de Catalunya* (2010FI_B 00521).

JOHN ROBERTO
Barcelona, 2015

RESUMEN

Esta tesis trata sobre el análisis de la polaridad en comentarios sobre productos, más exactamente, sobre la clasificación de comentarios como positivos o negativos a partir del uso de información lingüística. En la tesis presento un enfoque al análisis de la polaridad basado en el género discursivo de los comentarios. Según este enfoque, primero se identifican los segmentos que caracterizan el género discursivo de los comentarios y, posteriormente, se evalúa la utilidad que cada tipo de segmento tiene para determinar la polaridad de los comentarios.

La tesis se divide en dos partes. En la primera parte, caracterizo los comentarios como un género mediante el análisis de su estructura discursiva y su registro lingüístico. Sobre la base de ambos análisis postulo que los comentarios se componen de tres tipos principales de segmentos: valorativo, narrativo y descriptivo. En la segunda parte de la tesis, utilizo estos segmentos para calcular la polaridad de los comentarios. La hipótesis de partida es que no todos los segmentos que forman parte del género discursivo de los comentarios contribuyen de la misma manera a expresar la polaridad.

Para validar esta hipótesis evalúo tres métodos complementarios que tienen como objetivo detectar y determinar de forma automática la utilidad que tienen los tipos de segmentos para predecir la polaridad de los comentarios. El primer método utiliza información léxica y morfosintáctica para identificar el tipo de segmento que expresa mejor la polaridad del comentario. El segundo método analiza la función que desempeñan las secuencias narrativas en el cálculo de la polaridad. El tercer método se basa en el cálculo de la complejidad sintáctica para identificar y eliminar las oraciones que tienen una polaridad opuesta a la del comentario (oraciones asimétricas) como paso previo a la identificación de los comentarios positivos y negativos.

La conclusión principal que se desprende de estos análisis es que existe una relación directa entre el tipo de segmento y la polaridad expresada en el comentario: los usuarios suelen emplear de manera diferente los segmentos según se trate de un comentario positivo o uno negativo. Estas diferencias en el uso de los segmentos me ha llevado a plantear la existencia de dos (sub)géneros discursivos asociados a la expresión de opiniones sobre productos en la Web: el (sub)género de los comentarios positivos y el (sub)género de los comentarios negativos.

ABSTRACT

This thesis is about polarity analysis of reviews, that is, classifying reviews as either positive or negative based on linguistic evidence. I describe a genre-based approach for the polarity analysis of customer reviews. Genre is characterized by a schematic structure of the discourse composed of different types of stages, each one with a goal-oriented function. This approach to polarity analysis, first, distinguishes stages in the genre of reviews and, subsequently, evaluates the usefulness of each type of stage in the determination of the polarity of the entire review.

The thesis is broadly divided into two parts. In the first part, I characterize customer reviews as a discursive genre by analyzing both their structure and their linguistic register. Based on these analysis, I postulate that customer reviews are composed of three main types of stages: valorative, narrative and descriptive. In the second part of the thesis, I determine the usefulness of the different type of stages for the classification of positive and negative reviews. The rationale behind our approach is the assumption that within the specific genre of customer reviews, not all parts of a text contribute equally to expressing the underlying sentiment.

In order to validate this hypothesis, I evaluate three alternative methods used to automatically detect and determine the usefulness of each type of stage in the detection of the polarity of the entire review. The first method applies lexical and morphosyntactic information to identify the type of stage that best expresses the polarity of the whole review. The second method analyzes the role played by narrative chains in determining the polarity of reviews. The third method is based on the measurement of syntactic complexity to detect and remove descriptive sentences with the opposite polarity to that of the entire document (asymmetric sentences) as a previous step to identify positive and negative reviews.

The main conclusion that has been drawn is that there is a relationship between the types of stages and the polarity expressed in the review: users often employ stages differently according to whether its polarity is positive or negative. These differences in use of stages leads me to the conclusion that there are two (sub)genres, rather than one, for the expression of opinions on the Web: the (sub)genre of positive reviews and the (sub)genre of negative reviews.

ÍNDICE GENERAL

ÍNDICE DE CUADROS

ÍNDICE DE FIGURAS

1 INTRODUCCIÓN

Resumen

En este capítulo presento el problema de la detección de la polaridad en los comentarios sobre productos junto con mi propuesta para su tratamiento. Esta propuesta incluye los objetivos de la investigación, la hipótesis de trabajo y la descripción del procedimiento que aplico para su verificación.

1.1 DEFINICIÓN DEL ÁMBITO Y DEL PROBLEMA

Esta tesis trata el problema de la detección automática de la polaridad en comentarios sobre productos. Los comentarios sobre productos (*customer reviews*) constituyen un tipo de texto asociado a la expresión de opiniones en el marco de la Web social o colaborativa. Determinar si un comentario expresa una opinión positiva o negativa a partir de su contenido es una tarea de clasificación binaria de textos que se lleva a cabo desde el Análisis de la Polaridad. Las aproximaciones que existen para afrontar esta tarea suelen recurrir a información lingüística de diversa índole: morfológica, léxica, sintáctica y discursiva. La información discursiva se refiere a la estructura organizativa del texto a partir de las relaciones que se establecen entre sus diferentes elementos. El análisis de la polaridad mediante información discursiva está orientado a identificar y seleccionar las partes del texto de opinión que mejor expresen la polaridad. Ésta será la aproximación que seguiré en esta investigación.

Mi propuesta para detectar de forma automática la polaridad en los comentarios sobre productos se centra en utilizar la estructura discursiva de esta clase de textos de opinión para predecir su polaridad. Para ello identifico y selecciono los tipos de segmentos que caracterizan el género discursivo de los comentarios y los aplico para el cálculo de la polaridad. Mediante esta investigación he podido caracterizar los comentarios como un género discursivo autónomo y he determinado la función que cada uno de sus segmentos desempeña en la asignación de la polaridad. Este conocimiento ayudará a entender cómo los usuarios expresan las opiniones en lenguaje natural y facilitará la detección de la polaridad en diferentes dominios.

1.1.1 Conceptos clave

Esta investigación gira en torno a tres conceptos principales que son el COMENTARIO, la POLARIDAD y la ESTRUCTURA DE GÉNERO.

COMENTARIO: es un evento comunicativo que expresa una opinión, es decir, un sentimiento o una percepción acerca de determinada entidad o evento [B. Liu, 2008]. Frente a los enunciados objetivos («la película dura 90 minutos»), las opiniones buscan comunicar un estado emocional (*private state*) mediante el empleo de expresiones lingüísticas que expre-

san, por ejemplo, decepción («me esperaba algo más de esa película»), ira («¡¡vaya bodrio de peli!!, lástima el dinero de la entrada») o tristeza («una película para llorar»).

POLARIDAD: es un concepto que se usa para agrupar los sentimientos o estados emocionales en dos grandes clases[1]: positivos (alegría, placer, etc.) y negativos (disgusto, miedo, etc.). Por lo tanto, según el tipo de sentimiento que se relacione con la entidad o el evento valorado, una opinión puede ser positiva («una película *impresionante*») o negativa («la película es un *coñazo*»). De la misma manera, un comentario[2] puede ser positivo (👍) o negativo (👎) en función de la polaridad de las opiniones individuales que predominen en el texto, tal y como se aprecia en el Cuadro 1.

ASQUEROSO ★ ★ ★ ★ ★

Viejo, cutre, caro, sucio. Encima el trato es pésimo y te cobran por adelantado al llegar. No hay aire acondicionado, dicen que hay wi-fi y no funciona. *La ubicación es lo único decente, ya que queda a pocos metros de la playa y en pleno centro de Salou...* **Lo que por el otro lado quiere decir (puesto que hay que dormir con la ventana abierta por el calor) ruido hasta las 7 de la mañana. En una palabra: asqueroso.**

Cuadro 1: Ejemplo de un comentario con valoración negativa sobre un hotel (extraído de la web de TripAdvisor: `http://www.tripadvisor.es/`). Las opiniones negativas aparecen en negrita y las positivas en cursivas.

ESTRUCTURA DE GÉNERO: dentro de la Lingüística Sistémico Funcional la estructura de género o estructura génerica (*Generic Structure*) se define como el conjunto de elementos estructurales («partes», «instancias» o «segmentos») que un texto manifiesta como una propiedad específica de su género discursivo [Halliday, 1978]. Estos elementos estructurales, al tiempo que persiguen objetivos específicos, contribuyen a desarrollar una determinada estructura esquemática [Martin, 1992]. Halliday y Hasan [1985] abordan la estructura de género como un conjunto de estructuras posibles[3], jerarquizadas y ordenadas que están asociadas a un registro lingüístico particular.

1.1.2 Ámbito de trabajo

La detección automática de la polaridad en textos de opinión es un problema que involucra en diferente medida teorías, recursos y técnicas procedentes de varios campos de investigación. Aunque la Minería de Opiniones es el marco teórico general desde donde se aborda este problema, esta disciplina se apoya en otras disciplinas como son el Aprendizaje Automático, la Lingüística y el Procesamiento del Lenguaje Natural.

1 El RAE define la polaridad como la «*condición de lo que tiene propiedades o potencias opuestas, en partes o direcciones contrarias, como los polos*». Según el diccionario Collins, la polaridad es el «*estado de tener o expresar dos tendencias diametralmente opuestas, opiniones, etc.*».

2 Todos los comentarios que aparecen en esta tesis son versiones originales y no editadas de textos extraídos de Internet o de los corpus compilados para los experimentos.

3 El «potencial de estructura de género» [Halliday y Hasan, 1985] especifica los elementos estructurales obligatorios y optativos, junto con el orden opcional y obligatorio, y la mención de elementos de posible iteración que forman parte de un género discursivo determinado.

MINERÍA DE OPINIONES (MO) es el conjunto de aplicaciones y de técnicas que tienen como objetivo la extracción de información subjetiva a partir de contenidos generados por los usuarios. La tarea principal de la Minería de Opiniones es la detección de la polaridad, es decir, determinar la actitud (positiva o negativa) del hablante con respecto a un tema concreto. La tarea de asignar polaridad se puede llevar a cabo sobre las palabras aisladas, las oraciones (o partes de ellas) o sobre un documento completo. Dentro de la MO se suele usar la expresión ANÁLISIS DE LA POLARIDAD para referirse, exclusivamente, a la tarea de asignar polaridad al documento completo.

APRENDIZAJE AUTOMÁTICO (AA) es el estudio de métodos computacionales que pueden aprender y mejorar a partir de la experiencia. El objetivo del aprendizaje automático es «construir algoritmos que permitan obtener una descripción del concepto que subyace a un conjunto de observaciones o ejemplos de aprendizaje. Esta descripción debe ser coherente con el conjunto de observaciones y debe permitir predecir futuras observaciones del mismo problema» [Màrquez, 2002, p. 137]. El AA aplicado a la MO suele tratar el análisis de la polaridad como un problema de clasificación binaria entre las clases positiva y negativa.

LINGÜÍSTICA Y PROCESAMIENTO DEL LENGUAJE NATURAL (PLN). Si bien la lingüística tiene por objeto el estudio científico del lenguaje natural, la Web y la digitalización han implicado el surgimiento de nuevos medios de comunicación («*social media*») a los que esta disciplina ha de hacer frente. La espontaneidad e inmediatez de la comunicación que se produce en estos medios ha hecho posible que la comunicación escrita se enriquezca con nuevos **registros** y **nuevos géneros** discursivos, tal es el caso de los comentarios sobre productos. El tratamiento computacional de estos textos es una de las tareas del PLN. El PLN aplicado a la MO busca, entre otras cosas, proponer modelos computacionalmente efectivos del lenguaje humano que permita a las máquinas diferenciar una opinión positiva de una negativa. El papel de la lingüística es proporcionar el conocimiento necesario, en forma de rasgos o características (léxicas, morfosintácticas, semánticas o discursivas), para detectar la polaridad de los comentarios.

1.1.3 Resumen de la propuesta

Muchos investigadores en el campo de la Minería de Opiniones han abogado por el empleo de la información discursiva para el análisis de la polaridad de los textos (*discourse-based approaches*). Por ejemplo, Webber y col. [2011] aseguran que «la investigación sobre la estructura del discurso tiene un potencial considerable para contribuir al análisis de las opiniones, lo cual debería motivar intentos novedosos para integrar ambos campos de investigación». El análisis de la estructura del discurso permite entender los textos como una unidad compleja compuesta de diferentes segmentos que desempeñan una función específica en la expresión de los significados. Por lo tanto, el estudio de las relaciones entre los diferentes segmentos que integran los comentarios puede ayudar a establecer restricciones importantes sobre la manera en que se expresa normalmente la polaridad en esta clase de textos de opinión.

La propuesta que planteo en esta tesis consiste en aplicar la estructura del discurso al análisis de la polaridad de los comentarios sobre productos. Esta propuesta implica (i) identificar la estructura discursiva propia de los comentarios sobre productos, es decir, establecer los segmentos que los componen; y (ii) determinar la función que desempeña cada uno de estos segmentos en la expresión de la polaridad. Por lo tanto, el procedimiento que establezco para desarrollar la propuesta consta de dos etapas (ver la Figura 1):

PRIMERA ETAPA

La primera etapa consiste en la **caracterización de los comentarios como un género discursivo**. Esta etapa busca seleccionar los principales tipos de segmentos que hacen posible identificar los comentarios sobre productos como una clase específica de texto de opinión. Para ello:

- *Analizo* la estructura discursiva de los comentarios con el fin de determinar si comparten el mismo «propósito comunicativo» empleando, para ello, información léxica y morfosintáctica.
- *Analizo* el registro lingüístico[4] de los comentarios con el fin de determinar si comparten la misma «situación comunicativa».

El resultado de esta etapa, tal y como se puede ver en la Figura 1, es la selección de los tipos principales de segmentos que conforman los comentarios sobre productos.

SEGUNDA ETAPA

La segunda etapa consiste en el **análisis de la polaridad de los comentarios**. Esta tarea busca mejorar la detección de la polaridad de los comentarios mediante el tratamiento computacional de `cada uno` de los tipos de segmentos discursivos que lo componen. Para ello:

- *Clasifico* de forma automática cada uno de los tipos de segmentos que componen los comentarios basándome en diferentes características lingüísticas.
- *Calculo* la polaridad de los comentarios basándome en las propiedades discursivas y lingüísticas de cada tipo de segmento.

El resultado de esta etapa (ver la Figura 1) es la clasificación de los comentarios según su polaridad y unos niveles de precisión que sirven para determinar la función que desempeña cada tipo de segmento en la expresión de la polaridad del comentario.

La Figura 1 presenta las dos etapas propuestas para el análisis de la polaridad de los comentarios sobre productos. Cada etapa de la propuesta se presenta en un capítulo independiente: la primera etapa en el Capítulo 4 y la segunda etapa en el Capítulo 5. La ejecución de ambas etapas comporta el análisis lingüístico de los comentarios mediante el uso de corpus, analizadores (*parsers*), etiquetadores (*taggers*) y léxicos, entre otros recursos y herramientas. La información léxica, morfológica y sintáctica así obtenida, se usa como atributos de aprendizaje para entrenar múltiples clasificadores con los cuales abordar las tareas de caracterización del género discursivo y del análisis de la polaridad. En la primera etapa se analizan los comentarios desde el punto de vista estructural para seleccionar los segmentos que caracterizan este género discursivo. En la segunda etapa se evaluan tres métodos para clasificar de manera automática estos segmentos y aplicarlos al cálculo

4 Se denominan registro lingüístico al «nivel o modalidad expresiva» [Martínez de Sousa, 1995, p. 300] que selecciona el hablante con la finalidad de adaptarse a una determinada situación comunicativa.

de la polaridad del comentario completo: el primer método utiliza un conjunto de rasgos morfosintácticos para la detección y caracterización de los tipos de segmentos; el segundo método analiza la función que desempeñan las secuencias narrativas en el cálculo de la polaridad; y el tercer método se basa en el cálculo de la complejidad sintáctica para identificar y eliminar las oraciones que tienen una polaridad opuesta a la del comentario (oraciones asimétricas) como paso previo a la identificación de los comentarios positivos y negativos. A partir de los resultados se determina la función que desempeñan los segmentos en la expresión de la polaridad.

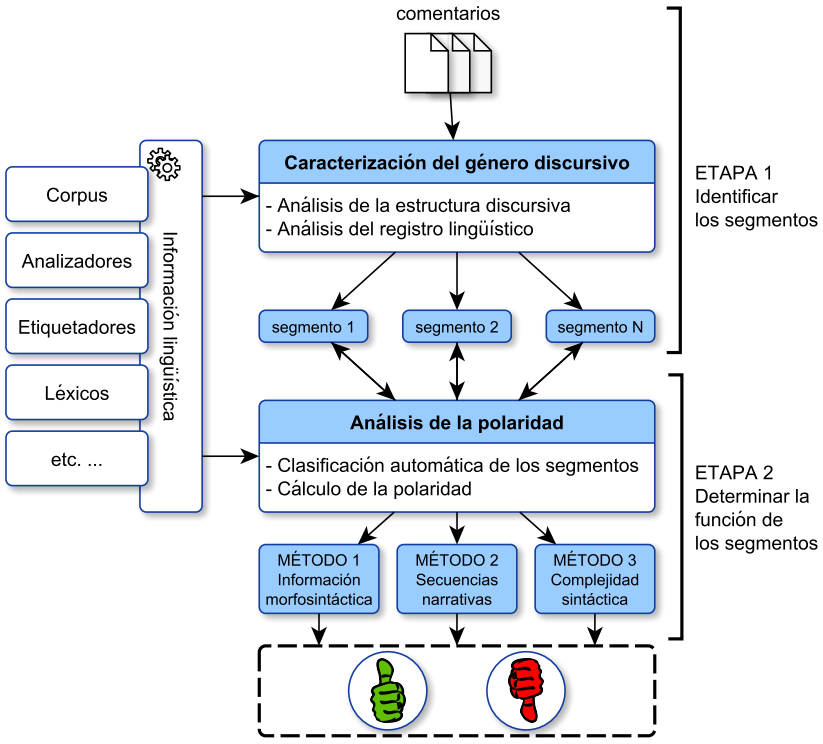


Figura 1: Esquema general para el análisis de la polaridad en comentarios sobre productos.

A continuación expongo los objetivos y las hipótesis de trabajo que guían el desarrollo de esta propuesta.

1.2 OBJETIVOS

El objetivo general de esta tesis es proponer métodos alternativos basados en la estructura del discurso para la detección automática de la polaridad en comentarios sobre productos. Este objetivo general se desglosa en dos objetivos específicos que están orientados a introducir mejoras cualitativas y cuantitativas en el análisis de la polaridad de esta clase de textos de opinión:

- Comprender los mecanismos discursivos que determinan la expresión de la polaridad en los comentarios sobre productos.
- Mejorar la detección de la polaridad en comentarios sobre productos mediante el tratamiento de su estructura discursiva.

Adicionalmente, el análisis de la polaridad de los comentarios basada en el tratamiento de su estructura discursiva da lugar al planteamiento de una serie de objetivos relacionados con la segmentación de los textos de opinión:

- Determinar si los comentarios sobre productos constituyen un género discursivo propio, caracterizado por la presencia de una estructura discursiva estable y por su empleo en situaciones comunicativas específicas.
- Identificar los tipos principales de segmentos que conforman los comentarios sobre productos y determinar la función que cumplen en la expresión de la polaridad.
- Plantear aproximaciones alternativas al tratamiento de la polaridad de los comentarios sobre productos que tengan en cuenta el propósito comunicativo de cada tipo de segmento que los componen.
- Proponer métodos para el tratamiento computacional de cada uno de esos tipos de segmentos.

Finalmente, la realización de esta tesis ha tenido también como objetivo:

- Generar herramientas y recursos para el tratamiento de la polaridad de los comentarios.

1.3 HIPÓTESIS

La hipótesis de trabajo principal de esta tesis es que el comentario constituye una unidad discursiva conformada por diferentes tipos de segmentos, cada uno de los cuales contribuye de forma diferente y con una intensidad específica en la expresión de la polaridad. Esta hipótesis se sustenta en las siguientes asunciones:

- **Asumo que el acto de opinar tiene lugar en el discurso.** La mayoría de las teorías lingüísticas contemporáneas tienen el discurso (no la oración) como unidad generadora de significado. En este sentido, Isenberg [1976] asegura que «cuando se produce una comunicación entre seres humanos es en forma de textos»[5]. Tres décadas después, desde el Marco Común Europeo de Referencia para las lenguas se afirma que «no puede haber un acto de comunicación por medio de la lengua sin un texto»[6] [de Europa, 2002].
- **Asumo que los géneros discursivos constituyen formas estructuralmente estables de representar los significados.** Según la Teoría de la Relevancia[7] [Sperber y Wilson, 1986; Wilson y Sperber, 2004] un enunciado es relevante cuando produce el máximo efecto con el menor esfuerzo cognitivo [Rauen, 2010, p. 28]. Para ello, los hablantes recurren a estructuras prototípicas o estandarizadas que resultan familiares al oyente. De acuerdo con Rauen [2009], los géneros textuales o discursivos desempeñan una función estandarizadora, facilitadora y transparente de acceso a los significados que tiene lugar a nivel del discurso.

5 Aquí el texto es entendido como un producto del discurso.
6 Ver nota 5 a pie de página.
7 La Teoría de la Relevancia es una teoría sobre la comunicación que explica el modo en que los hablantes interpretan los enunciados a partir del uso de inferencias.

- **Asumo que los comentarios comparten un mismo propósito comunicativo.** El acto de opinar sobre un producto en un foro, blog o un sitio web especializado, es un evento comunicativo reconocido social y culturalmente que forma parte de los medios de comunicación social para el consumo (*Consumer Generated Media*). El propósito de estos medios es facilitar la toma de decisiones en el ámbito comercial permitiendo a los usuarios la publicación y el acceso a las opiniones sobre productos.

1.4 ESTRUCTURA DE LA TESIS

El presente documento se estructura en seis capítulos que sintetizo a continuación.

Capítulo 1 Introducción

En este capítulo presento el problema de la detección de la polaridad en los comentarios sobre productos junto con los objetivos y la hipótesis de trabajo.

Capítulo 2 Antecedentes y estado del arte

En este capítulo efectúo una revisión de los antecedentes y del estado del arte relacionados con el análisis de la polaridad de comentarios. El objetivo de esta revisión es describir las diferentes propuestas que existen para representar los comentarios como un vector de características: bolsa de palabras (Sección 2.3.1), *N-gramas* (Sección 2.3.2), vectores sintácticos (Sección 2.3.3) y vectores de segmentos (Sección 2.3.4). Finalmente, en la Sección 2.4 presento las conclusiones del capítulo.

Capítulo 3 Recursos y herramientas

En este capítulo describo las herramientas y recursos que he aplicado en esta tesis. El objetivo de esta descripción es ofrecer una visión del procedimiento que he seguido para el tratamiento computacional de los comentarios. Este procedimiento consiste en tres etapas: compilación de los corpus de comentarios (Sección 3.2), procesamiento lingüístico de los comentarios (Sección 3.3) y clasificación de la polaridad de los comentarios (Sección 3.4). Finalmente, en la Sección 3.5 presento las conclusiones del capítulo.

Capítulo 4 Caracterización del género discursivo de los comentarios

En este capítulo caracterizo los comentarios como un género discursivo independiente. El objetivo de esta caracterización es vincular la expresión de polaridad al uso de determinados tipos de segmentos propios del género. En la primera parte del capítulo (Sección 4.2) presento los segmentos basándome en el reconocimiento de su propósito comunicativo y analizo su distribución en un corpus real. En la segunda parte del capítulo (Sección 4.3), presento varios experimentos orientados a identificar regularidades léxicas y morfosintácticas que permitan vincular esta clase de textos a una situación comunicativa específica (registro lingüístico). Finalmente, en la Sección 4.4 presento las conclusiones del capítulo.

Capítulo 5 Análisis de la polaridad de los comentarios

En este capítulo analizo la polaridad de los comentarios basándome en el tratamiento de su estructura discursiva. El objetivo de este análisis es identificar la función que desempeñan los diferentes tipos de segmento en el cálculo de la polaridad de los comentarios. En la Sección 5.2 presento los experimentos destinados a identificar de manera automática los segmentos discursivos que componen los comentarios. En las secciones 5.3 a 5.5 se proponen tres métodos alternativos para calcular la polaridad a partir de los segmentos discursivos. En la Sección 5.6 establezco la función que desempeña cada segmento en la expresión de la polaridad. Finalmente, en la Sección 5.7 presento las conclusiones del capítulo.

Capítulo 6 Conclusiones, contribuciones y trabajo futuro

En este capítulo recojo las conclusiones globales de este trabajo, las aportaciones del mismo y las líneas de trabajo futuro.

2 ANTECEDENTES Y ESTADO DEL ARTE

Resumen

En este capítulo efectúo una revisión de los antecedentes y del estado del arte en análisis de la polaridad de comentarios sobre productos. En los antecedentes defino las principales aproximaciones que existen para el tratamiento de la polaridad. La revisión del estado del arte está orientada a describir los diferentes tipos de información que se utilizan para crear los vectores con los que se suelen representar los textos de opinión: bolsa de palabras, Ngramas, características sintácticas y discursivas.

2.1 INTRODUCCIÓN

Gran parte de los contenidos en lenguaje natural generados en la web tienen un carácter subjetivo. Cada día se producen millones de mensajes orientados a opinar sobre productos a través de los servicios de microblogging (ej. Twitter[1]), redes sociales (ej. Facebook[2]), portales de opinión genéricos (ej. Ciao[3]) o especializados (ej. TripAdvisor[4]), por no mencionar las entradas de algunos blogs y los hilos de discusión en los foros públicos o privados.

Todos estos servicios desempeñan un papel informativo y muchas veces fundamental para mejorar la toma de decisiones. Tanto es así que las compañías y administraciones públicas se están interesando cada vez más por monitorizar las opiniones aparecidas en Internet acerca de los productos y servicios que ofrecen al público. Dicha monitorización la suele efectuar un *community manager* o responsable de comunidad de internet quien actúa como auditor de la marca en los medios sociales. No obstante, dado el elevado volumen de opiniones que se están produciendo, han surgido empresas dedicadas a desarrollar y aplicar software especializado para la gestión automática e inteligente de esta información.

Este marco tecnológico y lingüístico ha propiciado el surgimiento de una nueva disciplina: la Minería de Opiniones. Esta disciplina consiste en un conjunto de aplicaciones y de técnicas que tienen por objetivo la extracción de información subjetiva a partir de contenidos generados por los usuarios de la Web (blogs, tweets y comentarios, entre otros). La Minería de Opiniones está relacionada con la Minería de Textos (MT) y al Análisis de los Sentimientos (AS), pero se diferencia de éstas en la naturaleza de la información que analiza [Vinodhini y Chandrasekaran, 2012]. La Minería de Textos está enfocada al análisis de textos objetivos, es decir, aquellos en los que se narran eventos o se describen hechos (*facts*) de la realidad. El Análisis de los Sentimientos, por su parte, analiza la manera en que los «estados personales» (*private states*) se expresan en los textos subjetivos, es decir, en los textos

1 https://twitter.com
2 https://www.facebook.com/
3 http://www.ciao.com/
4 http://www.tripadvisor.es/

que comunican el punto de vista de su autor. Siguiendo a Cambria y col. [2013], la tarea principal del Análisis de los Sentimientos es el reconocimiento de las emociones (alegría, pena, ira, etc.). La Minería de Opiniones en cambio, tiene por objetivo la detección de la polaridad (positiva y negativa): «Aunque comúnmente se utilizan de forma indistinta para denotar el mismo campo de estudio, la minería de opiniones y el análisis de sentimiento se centran en la detección de la polaridad y el reconocimiento de emociones, respectivamente» [Cambria y col., 2013, p. 15].

En este capítulo reviso los antecedentes y el estado del arte en análisis de la polaridad de comentarios sobre productos. En la Sección 2.2 presento los primeros trabajos orientados al análisis de la polaridad y las principales técnicas mediante las cuales se lleva a cabo esta tarea. En la Sección 2.3 detallo las diferentes propuestas que existen para representar los comentarios como un vector de características que mejor expresen la polaridad. Finalmente, en las conclusiones, relaciono mi propuesta con los enfoques existentes.

2.2 ANÁLISIS DE LA POLARIDAD: ANTECEDENTES

Los primeros trabajos dedicados al análisis de la polaridad corresponden a Tong [2001], Turney [2002], Pang, L. Lee y Vaithyanathan [2002], Dave y col. [2003], Grefenstette y col. [2004] y Cui y col. [2006]. Tong [2001] es pionero en el área. Este autor llevó a cabo un análisis automático de textos de opinión con el fin de predecir la opinión de los consumidores sobre determinados productos. Uno de los trabajos más citados en el área es el de Turney [2002], quien propuso detectar la polaridad de comentarios estimando la orientación semántica de algunas de las palabras que componen los textos. No obstante, fueron Dave y col. [2003] quienes utilizaron el término «minería de opiniones» por primera vez para describir una herramienta capaz de «procesar un conjunto de resultados de búsqueda para un elemento determinado, generar una lista de los atributos del producto y agregar las opiniones acerca de cada uno de ellos» (ej. precio:alto, calidad:baja). Estos trabajos motivaron el surgimiento de nuevos enfoques y técnicas para aproximarse al problema.

En general, se han utilizado dos aproximaciones para abordar el problema de la detección de la polaridad de los textos de opinión: la aproximación basada en léxicos (análisis no supervisado) y la aproximación basada en aprendizaje automático (análisis supervisado) [Pang y L. Lee, 2008, p. 27; Taboada, Brooke, Tofiloski y col., 2011, p. 268; Fernández y col., 2014, p. 19]. La aproximación basada en léxicos, también denominada semántica (*non-supervised semantic-based method*), se fundamenta en el uso de recursos externos, principalmente léxicos en los que se declara la polaridad de base de algunas palabras (o grupos de palabras). La segunda aproximación utiliza técnicas de aprendizaje automático. Esta aproximació requiere un corpus que contenga textos previamente etiquetados con su polaridad (conjunto de entrenamiento) para crear un clasificador que sea capaz de etiquetar nuevos textos. A diferencia de la aproximación semántica, los métodos de aprendizaje automático proporcionan un rendimiento razonablemente alto en el dominio en que han sido entrenados [Rothfels y Tibshirani, 2010, p. 2]. No obstante, su rendimiento cae cuando se utilizan en un dominio diferente al de entrenamiento [Lin y col., 2012; Pang y L. Lee, 2008; Tan y col., 2009]. La eficiencia también puede verse afectada si el número de características es muy grande o los datos para el entrenamiento son escasos [Aue y Gamon, 2005].

El trabajo de Turney [2002] fue el primero en establecer una metodología centrada en el uso de métodos no supervisados para clasificar documentos atendiendo a su polaridad. En su artículo, el autor presenta un algoritmo de aprendizaje no supervisado para calificar un comentario como positivo (*thumbs up*) o negativo (*thumbs down*) que desarrolla en tres pasos: extracción de palabras o frases que expresen orientación semántica, determinación de su polaridad y cálculo de la polaridad del texto combinando las polaridades individuales de las palabras o frases presentes en el texto. La principal ventaja de los métodos no supervisados es su facilidad para ser aplicados sobre diferentes dominios de conocimiento aunque, de otro lado, sus resultados tienden a no superar la calidad de los métodos supervisados.

Pang, L. Lee y Vaithyanathan [2002] fueron los primeros en utilizar métodos supervisados para clasificar documentos atendiendo a su polaridad. Estos investigadores analizaron, desde el punto de vista del Aprendizaje Automático, si era viable tratar la polaridad como un caso especial de clasificación de textos según el tema. Para validar su hipótesis aplicaron tres algoritmos de clasificación (`Bayes`, `Maximum Entropy` y `Support Vector Machines`) sobre un corpus de más de 2000 comentarios sobre películas obtenidas de *Internet Movie Database*[5] (IMDb). Las conclusiones extraídas de los experimentos demostraron que no era posible obtener niveles de precisión sobre el problema de la clasificación de la polaridad en los textos comparables a los reportados para la categorización basada en el tema del documento utilizando los mismos algoritmos. En este sentido, la principal dificultad de los métodos supervisados aplicados al análisis de la polaridad ha sido y sigue siendo la forma de representar los datos de entrada para generar el vector de características asociado con el par de clases objetivo (positivo / negativo).

En la siguiente sección describo las diferentes propuestas que existen para generar los vectores con las características que se emplean para representar los textos de opinión.

2.3 RASGOS USADOS PARA LA REPRESENTACIÓN VECTORIAL DE LOS COMENTARIOS

En Aprendizaje Automático, la precisión de los algoritmos de clasificación depende en gran medida de la selección del tipo de rasgos que se usen para describir el dominio, conjuntamente con la forma de representar tales rasgos (vectores, matrices, árboles, etc.). En Minería de Opiniones el tratamiento de la polaridad se suele afrontar mediante el uso de vectores conformados, básicamente, por cuatro clases de rasgos: palabras, n-gramas, dependencias sintácticas y relaciones discursivas. En esta sección describo en qué consiste cada una de estas clases de información y cómo se usan para detectar la polaridad en los comentarios.

2.3.1 Vector de bolsa de palabras

La bolsa de palabras (*bag-of-words*, BoW) [Manning y Schutze, 1999] es la representación más tradicional y simple que se puede obtener de los comentarios. Este método utiliza las palabras de los comentarios como los elementos del vector de términos sin tener en cuenta el orden en el que aparecen. Sobre esta representación es posible entrenar diferentes clasificadores (prin-

5 `http://www.imdb.com/`

cipalmente `Support Vector Machines` y `Naive Bayes`). La presencia y/o frecuencia de las palabras como variables en los vectores de aprendizaje ha sido una de las aproximaciones más utilizadas en la tarea de clasificación de la polaridad: Pang, L. Lee y Vaithyanathan [2002], Dave y col. [2003], Mullen y Collier [2004], Aue y Gamon [2005].

Algunas de las limitaciones que tiene el uso de BoW son: su alta dimensionalidad, ya que requiere un gran número de recursos computacionales para llevar a cabo la clasificación; su elevado nivel de dispersión de la información, lo que afecta el rendimiento de la mayoría de los algoritmos de Aprendizaje Automático; su incapacidad para captar correctamente el contexto, y no preservar ningún tipo de relación entre las palabras y las clases.

Para reducir la alta dimensionalidad del modelo BoW se suele eliminar del vector de términos las palabras que más se utilizan en cada idioma. A este conjunto de palabras se le conoce como el conjunto de palabras vacías, palabras funcionales o *stopwords*. Otras técnicas de reducción son el *stemming* y la *tokenization*. Estas técnicas están orientadas a reducir las variantes morfológicas de las palabras a su forma léxica común. Por último, aunque no menos importante, algunas representaciones basadas en BoW suelen incluir diferentes esquemas de pesado de las palabras, como puede ser `tf-idf`, con los que se intenta expresar el poder discriminatorio de un término en una colección de documentos.

Paltoglou y Thelwall [2010] y Martineau y Finin [2009] han trabajado para optimizar la bolsa de palabras en la detección de la polaridad. Según Paltoglou y Thelwall [2010], los esquemas de pesado tradicional se deben adaptar a las peculiaridades de los textos de opinión puesto que su funcionamiento se sustenta en la selección de las palabras menos frecuentes, y las palabras que expresan una emoción son muy frecuentes en los corpus de opinión. Por lo tanto, para reducir la penalización de palabras más frecuentes proponen optimizar el valor de `tf-idf` calculándolo en los documentos positivos y negativos por separado. De esta manera, se da relevancia a las palabras menos comunes en cada tipo de documento. Tanto Paltoglou y Thelwall [2010] como Martineau y Finin [2009] han comprobado cómo `Delta tf.idf`, un esquema de pesado más «general e intuitivo», mejora la precisión en tareas de clasificación de la polaridad. Ambos autores utilizan `Support Vector Machines` para el aprendizaje.

En general, las principales ventajas de utilizar el modelo BoW son su simplicidad y eficacia. Estas características lo convierten en un modelo ideal para ser usado como *baseline* en múltiples y variados estudios sobre la polaridad. En esta línea cabe destacar los trabajos de Tu y col. [2012], Mukherjee y Bhattacharyya [2012], Ghorbel y Jacot [2011], Martineau y Finin [2009], Na y Thet [2009], Aue y Gamon [2005], Pang, L. Lee y Vaithyanathan [2002], entre otros. Además, Pang, L. Lee y Vaithyanathan [2002] demostraron tempranamente que la representación vectorial más básica de los comentarios, como la fundamentada en la presencia y ausencia de palabras (representación binaria), muchas veces proporciona un mejor rendimiento que representaciones más complejas, como las sustentadas en la frecuencia de los términos. Pang, L. Lee y Vaithyanathan [2002] y Aue y Gamon [2005] reconocen que la bolsa de palabras con `Support Vector Machines` proporciona un *baseline* bastante robusto para problemas de clasificación de la polaridad de textos.

El Análisis Semántico Latente (*Latent Semantic Analysis, LSA*) se ha empleado como método para reducir la dimensionalidad del modelo de espacio vectorial y priorizar las dimensiones con mayor contenido de información [Xie y col., 2014; Yih y col., 2012]. El LSA es un modelo estadístico del uso de pala-

bras que permite comparar las similitudes semánticas entre dos fragmentos de información textual (vectores, $\vec{v}, \vec{w}$) pertenecientes a un mismo dominio de conocimiento. El LSA se suele aplicar para determinar la orientación semántica de los términos de los comentarios puesto que este modelo permite encontrar las palabras que se comportan de manera similar basándose en el análisis de sus contextos. Por ejemplo, Turney [2002] desarrolla un algoritmo que mide la «distancia» que hay entre los términos y un par de palabras semilla de las que se conoce su polaridad, por ejemplo, *excelent* (positiva) y *poor* (negativa). Esta distancia viene determinada por la co-ocurrencia entre el término y cada una de las semillas.

El LSA suele estar condicionado por las longitudes de los vectores de ambos términos, además no posee ningún tipo de análisis sobre el orden o el rol de las palabras [Botana y col., 2008]. Sobre este último aspecto, el rol, Yih y col. [2012] afirman que el LSA sólo considera las relaciones de sinonimia (palabras que tienen los mismos patrones de co-ocurrencia) pero no las de antonimia. Esta limitación incide negativamente en la determinación de la orientación semántica de los términos y, en general, en el cálculo de la polaridad de los comentarios. Yih y col. [2012] proponen una variante de LSA enfocada al tratamiento de la polaridad en la que también se incorporan al espacio vectorial las palabras que tienen significados opuestos. A. L. Maas y col. [2011] presentan un algoritmo que además de captar la similaridad semántica entre palabras, también capta su similaridad a nivel de polaridad: por ejemplo, las palabras *wonderful* y *amazing* no sólo están próximas semánticamente sino que ambas son marcadamente positivas. El algoritmo de A. L. Maas y col. fue evaluado exitosamente en una tarea de clasificación de la polaridad empleando tanto métodos supervisados como no supervisados. Xie y col. [2014] también proponen una variante del LSA que integra información morfosintáctica en el modelo (*Tag Sentiment Aspect Models*, *TSA*).

2.3.2 Vector de *N-gramas*

Si en lugar de tomar las palabras de forma individual se toman secuencias de n palabras, tenemos lo que se conocen como *n-gramas*: «demasiado-costoso», «muy-buen-producto», «no-me-agrada», etc. Además de ser un modelo de fácil representación y obtención, los *n-gramas* captan mejor la expresión de polaridad de un texto que el modelo BoW. Por ejemplo, muchas veces el término *bad* suele ser sustituido por el *bigrama not-good* en los comentarios negativos. Por este motivo, numerosos sistemas para el cálculo de la polaridad utilizan *n-gramas* para representar las características de aprendizaje.

A diferencia de los clasificadores que categorizan textos por la temática, los «clasificadores de la polaridad» se entrenan seleccionando *n-gramas* con expresiones positivas y negativas e incorporando al proceso la negación u otros modificadores de la polaridad o *valence shifters* [Polanyi y Zaenen, 2005]. Es el caso de S. Li y col. [2010], quienes presentan una aproximación para el tratamiento de la polaridad de comentarios sobre productos basada en la detección de los modificadores de la polaridad en cuatro dominios distintos (libros, DVDs, electrónica y electrodomésticos de cocina). Su método consiste en detectar y separar, mediante el uso de *n-gramas*, las oraciones que contienen modificadores de la polaridad (ej. «*I am **not** happy*») de las que no contienen modificadores (ej. «*I am disappointed*») como paso previo al análisis de la polaridad de los comentarios. Kouloumpis y col. [2011] también aplican *n-gramas* para incorporar la negación y los emoticonos al vector de

características. Como algoritmos de aprendizaje utilizan `AdaBoost` y `Support Vector Machines`.

Un procedimiento habitual es generar los *n-gramas* a partir de información morfosintáctica que se obtiene mediante lematizadores y etiquetadores morfológicos (*part-of-speech taggers*). Mediante un lematizador se obtiene la forma canónica (lema) de las palabras y mediante el etiquetador o analizador morfológico se obtiene su categoría gramatical. Uno de los objetivos de este procedimiento es conformar los *n-gramas* solo con aquellos tipos de palabras que presentan carga emocional como los adjetivos. En la primera parte de su trabajo, Na, Sui y col. [2004] demostraron que el rendimiento de un clasificador `Support Vector Machines` disminuía cuando, además de adjetivos, adverbios y verbos, se empleaban sustantivos. Adicionalmente, en la segunda parte de su trabajo, usaron la información morfosintáctica para detectar *n-gramas* que contenían expresiones negativas (ej. *I'd never regretted purchasing it*), mejorando los resultados de la detección de la polaridad de los comentarios. Ebrahim y col. [2012] emplean los adjetivos y adverbios para identificar las oraciones que expresan opinión y los nombres y sintagmas nominales para reconocer las partes de estas oraciones que tratan sobre alguna característica del producto valorado. Mediante WordNet[6] y siguiendo la idea de Turney [2002], los autores resuelven la orientación semántica de los adjetivos en las oraciones que contienen características muy frecuentes. Finalmente, entrenan un clasificador binario basado en `Support Vector Machines` a partir de los comentarios anotados con esta información y previamente etiquetados con la polaridad.

Pang, L. Lee y Vaithyanathan [2002] seleccionaron 16.165 *bigramas* que, agrupados en ocho configuraciones diferentes, utilizaron para entrenar tres algoritmos de aprendizaje: `Naive Bayes`, `Maximum Entropy` y `Support Vector Machines`. Los autores también tuvieron en cuenta la posición de los *bigramas* en los comentarios con el fin de asignarles más o menos peso discriminatorio. De manera similar, Turney [2002] propuso un algoritmo de aprendizaje no supervisado para clasificar como positivos o negativos los comentarios de varios dominios. La clasificación de los comentarios se realizó promediando la polaridad de *n-gramas* obtenidos mediante determinados patrones morfosintácticos: «*JJ/little NN/difference*», «*JJ/subtle NNS/nuances* », «*RB/very JJ/cavalier*». Su algoritmo consta de tres pasos: extracción de los patrones, cálculo de la orientación semántica de cada patrón y clasificación del comentario.

En la misma línea está el trabajo de Bakliwal y col. [2011] quienes fundamentan su algoritmo para clasificar la polaridad de comentarios en la combinación de *n-gramas* de palabras (ej. *good-product*) con *n-gramas* de categorías morfosintácticas (ej. *good_JJ-product_NN*). Sobre este espacio de características entrenan varios clasificadores: `Naive Bayes`, `Multi - Layer Perceptron` y `Support Vector Machines`. Amiri y Chua [2012] aplican *n-gramas* de categorías morfosintácticas e información semántica obtenida de WordNet para determinar la polaridad de comentarios en múltiples dominios. Garg [2013] aplica *n-gramas* sobre clasificadores `Naive Bayes` y `Support Vector Machines` al tratamiento de la polaridad de comentarios sobre películas en idioma Punjabi. También Saleh y col. [2011] utilizan `Support Vector Machines` con *unigramas*, *bigramas*, *trigramas* y varios esquemas de pesado para clasificar comentarios en diferentes dominios. Graovac y Pavlovic-Lazetic [2014]

6 `https://wordnet.princeton.edu/`

utilizan *n-gramas* de caracteres[7] (*byte-level n-gram*) para detectar la polaridad de comentarios con independencia del idioma y del dominio.

Finalmente, Lebret y Collobert [2015] presentan un estudio comparativo en el que analizan el rendimiento de varios clasificadores utilizando múltiples características entre las que se encuentran BoW y *n-grams*. Dentro de las técnicas que evalúan están *Latent Semantic Analysis* o *Latent Dirichlet Allocation* y el modelo *Skip-gram*. El uso de *skip-gram* [Guthrie y col., 2006] es una técnica para reducir la dispersión del vector de datos basada en la construcción de *n-gramas* de elementos no continuos. Por ejemplo, la oración *Insurgents killed in ongoing fighting* se compone de los siguientes *2-skip-bi-grams*: insurgents_killed, insurgents_in, insurgents_ongoing, killed_in, killed_ongoing, killed_fighting, in_ongoing, in_fighting, ongoing_fighting. También, Wang y col. [2014] y Fernández y col. [2014a] utilizan *skip-gramas* como unidad de información para entrenar clasificadores de la polaridad.

2.3.3 Vectores sintácticos

Dado que los *n-gramas* ignoran las relaciones estructurales que se establecen de manera discontinua entre diferentes elementos de la oración (ej. «**esto no es**, en mi opinión, **muy bueno**»), se ha incorporado el análisis sintáctico para introducir este tipo de información en los vectores de clasificación. Si bien, el uso de características basadas en el análisis sintáctico constituye una alternativa a los *n-gramas* para hacer generalizaciones sobre la estructura de los textos de opinión, la construcción de estos vectores suele exigir muchos recursos sin que ello resulte en un mejor rendimiento. En efecto, tal como sostienen Xu y col. [2014], Joshi y Penstein [2009], Pang [2006] y Dave y col. [2003], el uso de estos rasgos o características para la Minería de Opinión ha arrojado resultados contradictorios. Xu y col. [2014], por ejemplo, determinaron que la información semántica era más efectiva que la sintáctica para llevar a cabo tareas de minería de rasgos.

Este tipo de análisis puede realizarse empleando analizadores sintácticos globales (*deep parsers*) o parciales (*shallow parsers*). El objetivo de los analizadores sintácticos globales es abordar cualquier tipo de construcción mediante el análisis robusto de textos no restringidos. No obstante, debido a su complejidad, coste computacional y limitaciones de fiabilidad, los sistemas para el tratamiento de la polaridad suelen basarse en un análisis parcial o superficial de los comentarios. El análisis sintáctico superficial devuelve una representación aproximativa e incompleta de la estructura sintáctica ya que identifica los constituyentes de la oración sin llegar a especificar su estructura interna. Según Abney [1996], «el análisis parcial tiene como objetivo recuperar información sintáctica de forma eficiente y fiable a partir de texto no restringido, sacrificando la complejidad y profundidad del análisis global».

El análisis sintáctico en Minería de Opiniones está orientado, principalmente, a identificar las expresiones lingüísticas que permiten recuperar las características de un producto [Mukherjee y Bhattacharyya, 2012; Wu y col., 2009; zhang y col., 2009], así como las relaciones que se establecen entre tales expresiones, por ejemplo, de correferencia[8] («*I highly recommend the* ***Canon SD500*** *to anybody looking for a compact* ***camera***»). Ding y B. Liu [2010]

7 Por ejemplo, la oración *No pain, no gain!* se compone de los siguientes *3-gramas*: N o _; o _ p; _ p a; p a i; a i n; i n ,; n , _ ; , _ n; _ n o; n o _; o g; _ g a; g a i; a i n; i n !; (donde el guion bajo («_») representa un espacio en blanco).

8 «Las relaciones de correferencia [...] se establecen entre expresiones lingüísticas que se refieren a una misma persona, objeto o acontecimiento.» [Recasens, 2010]

han estudiado la resolución de la correferencia en textos de opinión, Kessler y Nicolov [2009] la correferencia y la meronimia[9] (un caso de meronimia citado por los autores: «*I also looked at the **Toyota Corolla**, but its **engine** seemed sluggish*»). Estas investigaciones trabajan a un nivel más detallado (*fine grained*) resolviendo la polaridad del sintagma y la oración para luego hacerla extensiva a todo el comentario. Un ejemplo representativo de esta aproximación la podemos encontrar en C. Zhang y col. [2009], donde los autores aplican dependencias sintácticas, algoritmos basados en reglas y aprendizaje automático (`Support Vector Machines`, `Naive Bayes` y `Decision Trees`) para determinan la polaridad de cada una de las oraciones en un conjunto de comentarios en chino y, a continuación, utilizan los resultados de este análisis para predecir la polaridad de los documentos. Evidentemente, también es posible usar la polaridad del comentario para predecir la polaridad de las oraciones o sintagmas individuales [Y. Zhang y col., 2014].

Otro grupo importante de los trabajos que aplican analizadores sintácticos están orientados a identificar los modificadores de la polaridad y las expresiones subjetivas (oraciones, sintagmas o constituyentes) que luego son clasificadas de acuerdo con su polaridad. Matsumoto y col. [2005] entrenan clasificadores basados en `Support Vector Machines` a partir de constituyentes sintácticos que son recurrentes en los comentarios sobre películas puesto que, según los autores, estos constituyentes suelen conformar expresiones subjetivas (ej. *the film, however, is all good*). Una aproximación similar es la de J. Liu y Seneff [2009], pero centrándose en los sintagmas compuestos por modificadores de la polaridad como las negaciones (`not(very(good))`). Joshi y Penstein [2009] aplican un *parser* de dependencias con el que extraen tripletas de la forma `relación_núcleo_modificador` (ej. `amod_camera_great`) que utiliza para generar el vector de características sobre un conjunto de comentarios. W. Zhang y col. [2009], por su parte, presentan SESS (*SElf-Supervised and Syntax-Based method*), un método enfocado a la detección de términos de polaridad mediante el tratamiento de oraciones coordinadas (ej. *big and nive*), condicionales (ej. *... even if you did find your room small*) y concesivas (ej. *great hotel, however the staff ...*). Buczynski y Wawer [2008] intentan suplir la falta de estudios en idiomas diferentes al inglés aplicando el análisis sintáctico superficial para la detección de la negación en comentarios sobre productos en polaco. Farra y col. [2010] presentan un trabajo similar pero para el árabe.

En la misma línea, Tu y col. [2012] aplican *kernels* (núcleos, máscaras o matrices) de convolución[10] para codificar diversas estructuras sintácticas presentes en los comentarios sin necesidad de establecer reglas explícitas. Vilares y col. [2013] describen una aproximación sintáctica para el tratamiento de la polaridad de comentarios en español basada en el análisis sintáctico de dependencias. La estructura sintáctica se emplea para tratar la intensificación, las oraciones subordinadas adversativas y la negación. Según los autores de este trabajo, una de las ventajas de esta aproximación es que permite contrarrestar parte de los efectos negativos originados por la «tendencia positiva del lenguaje humano», es decir, la necesidad de expresar en términos positivos las valoraciones negativas de los productos: «al expresar una opinión negativa, es frecuente utilizar negaciones de términos positivos en lugar de los correspondientes antónimos; "no barato" en vez de "caro" o "no bueno" en vez de "malo"».

9 La meronimia es la relación semántica entre una expresión lingüística que denota una parte y otra que denota el todo.

10 El *kernel* de convolución consiste en el tratamiento de una matriz por otra que se denomina «*kernel*» con el objeto de eliminar los elementos innecesarios.

2.3.4 Vectores de segmentos

Finalmente, comentaré algunas aproximaciones que tienen en cuenta la estructura del discurso para representar el espacio de características sobre el que se calcula la polaridad de los comentarios.

Desde la Minería de Opiniones son múltiples las voces que han expresado la necesidad de discriminar entre los varios tipos de segmentos que caracterizan los textos de opinión: Webber y col. [2012], Heerschop y col. [2011], Taboada, Brooke y Stede [2009], Bieler y col. [2007], Polanyi y Zaenen [2005], Pang y L. Lee [2004], Turney [2002], por poner algunos ejemplos. De acuerdo con Webber y col. [2012, p. 467-468], incluir la información sobre la estructura del discurso para identificar las partes de un discurso cuyas expresiones evaluativas son particularmente relevantes puede mejorar el rendimiento en el análisis de las opiniones. Esta necesidad de incorporar la información discursiva al análisis de los textos de opinión no es nueva. Pang y L. Lee [2004, p. 277] sostenían que «la detección de los segmentos subjetivos ayuda a reducir la extensión de los comentarios en fragmentos más breves que siguen conservando información sobre la polaridad a un nivel comparable a la del comentario completo».

Las aproximaciones al análisis de la polaridad a partir de la estructura del discurso se pueden dividir en dos grandes grupos: las basadas en *parsers* discursivos (*discourse parsing*) y las basadas en la segmentación (*discourse zoning*). El primer grupo está representado por dos teorías, la Teoría Representacional del Discurso Segmentado (*Segmented Discourse Representation Theory*, SDRT) [Lascarides y Asher, 2007] y la Teoría de la Estructura Retórica (*Rhetorical Structure Theory, RST*) [Mann y Thompson, 1988]. Las aproximaciones basadas en la segmentación son más heterogéneas, aunque suelen fundamentarse en el género discursivo de los textos. A continuación comento algunos de los trabajos adscritos a cada una de estas líneas de investigación.

Chardon y col. [2013] estudian la relación que hay entre la estructura del discurso, el género textual y la polaridad en textos de opinión utilizando la Teoría Representacional del Discurso Segmentado como marco de referencia. Los investigadores utilizan un segmentador discursivo para identificar las Unidades de Discurso Elementales (*Elementary Discourse Units*, EDUs) en un conjunto de comentarios sobre películas y noticias. A partir de la segmentación proponen tres estrategias para calcular la polaridad de los textos: bolsa de segmentos (*bag of segments*), análisis parcial del discurso (*partial discourse*) y análisis completo del discurso (*full discourse*). Los buenos resultados obtenidos mediante su aproximación confirman que la estructura del discurso desempeña un papel fundamental en la asignación de la polaridad de los comentarios.

Utilizando la Teoría de la Estructura Retórica como marco de referencia, Heerschop y col. [2011] investigan si la estructura discursiva aporta información útil para clasificar documentos según su polaridad. Los investigadores segmentan los comentarios en sus Unidades de Discurso Elementales y extraen las relaciones retóricas estructurales establecidas entre dichas unidades. A continuación, asignan pesos a las palabras según la posición que ocupen en el comentario o en la estructura jerárquica de los árboles retóricos (*RST trees*) y según el tipo de relación del que forme parte (contraste, explicación, elaboración, etc.). Heerschop y col. demuestran que la clasificación de los documentos según la polaridad mejora al considerar los diferentes tipos de relaciones retóricas que dan coherencia al texto.

No obstante, la Teoría de la Estructura Retórica ha sido cuestionada en varios aspectos. En Taboada y Mann [2006] y Reed y D. Long [1998] se señala como puntos débiles, que la RST constituye básicamente un modelo descriptivo antes que una herramienta para la generación y comprensión de textos, la falta de estudios y de recursos (corpus, por ejemplo) en idiomas distintos al inglés y el que las mismas relaciones retóricas se realicen de manera diferente en diferentes géneros discursivos. Adicionalmente, tanto la RST como la SDRT presentan serias limitaciones de carácter práctico como son el coste computacional y la complejidad de la tarea de segmentación.

Precisamente, debido a su elevado coste computacional, la RST se ha aplicado al análisis de la polaridad en escenarios específicos y reducidos. Por este motivo Chenlo y col. [2014] y Chenlo y col. [2013] estudian el rendimiento de la RST en un amplio conjunto de textos de opinión compuesto por blogs, comentarios sobre productos y artículos de opinión. Por razones de eficiencia, los autores sólo procesan con el *parser* discursivo aquellas oraciones que expresan opinión y que se han recuperado previamente mediante *OpinionFinder*[11]. El objetivo del uso de un *parser* discursivo es identificar, en cada una de estas oraciones, el fragmento que expresa la polaridad y que según la RST suele coincidir con el segmento nuclear. Por ejemplo, en una relación de contraste el núcleo y no el satélite de la relación expresa la polaridad: *[Although I like the characters,]*$_{\mathrm{sat}}$ ***[the book is horrible]***$_{\mathrm{nuc}}$.

En términos generales, para extraer las relaciones retóricas de los comentarios es necesario un *parser* discursivo robusto y eficiente, y los *parsers* actuales presentan serias limitaciones en ambos aspectos. Según Feng e Hirst [2014], el *parser* discursivo más eficiente es de Joty y col. [2013], con un nivel de rendimiento de sólo el 55.7 %. Por este motivo, en algunas ocasiones el análisis de la polaridad a nivel discursivo se realiza sin el uso de estas herramientas. En estos casos, la segmentación del comentario se efectúa seleccionando unidades formales básicas como los párrafos, secciones (ej. título, pros, contras, etc.) u oraciones. También es común seleccionar las partes del texto que se encuentren en una determinada posición (ej. principio, medio o final) o que cumplan con algunas propiedades (ej. oraciones subjetivas). Estas aproximaciones al tratamiento del discurso suelen requerir una fase previa al análisis de la polaridad enfocada a entrenar y evaluar el rendimiento del algoritmo que se propone para extraer las partes relevantes del comentario. Algunos de los primeros trabajos en Minería de Opiniones valoraron la posibilidad de segmentar los textos como una alternativa al tratamiento de su polaridad.

En efecto, por razones evidentes, desde sus inicios la Minería de Opiniones ha tenido claro que los segmentos evaluativos o valorativos son más relevantes que los descriptivos a la hora de identificar la polaridad de un texto. El primer paso en el algoritmo de Turney [2002] consiste en identificar los sintagmas que contienen adjetivos o adverbios. Una vez identificados y asignada su orientación semántica, se decide la polaridad del comentario. Pang y L. Lee [2004] también asumen que la polaridad se expresa mejor en los segmentos subjetivos por lo que presentan una aproximación para el análisis de la polaridad de los comentarios basada en la recuperación de las oraciones subjetivas (ej. «*This is a good movie*») y la consiguiente eliminación de las oraciones objetivas (ej. «*The protagonist tries to protect her good name*»). Según los autores, las oraciones objetivas no aportan información sobre la

11 *OpinionFinder* es un sistema que identifica oraciones subjetivas en gran variedad de textos. El sistema, desarrollado por la Universidad de Pittsburgh, se puede obtener gratuitamente desde su web: `http://mpqa.cs.pitt.edu/opinionfinder/`.

opinión del autor del comentario por lo que merman el rendimiento de los clasificadores de la polaridad.

Bieler y col. [2007] hacen uso del género discursivo –entendido como un tipo de texto con una estructura socialmente reconocible– para detectar la polaridad en textos de opinión. Los autores segmentan los comentarios sobre películas en párrafos formales (ej. director, argumento) y funcionales (descripciones y valoraciones), propios de este (sub)género de opinión, y emplean Aprendizaje Automático basado en BoW para clasificar dichas zonas. Taboada y Grieve [2004] y Taboada, Brooke y Stede [2009] también se valen del género discursivo para seleccionar las partes de los comentarios que mejor expresan una opinión. En efecto, siguiendo el trabajo de Bieler y col. [2007] y apoyándose en la hipótesis de que las opiniones tienden a expresarse en partes específicas de los comentarios, Taboada y Grieve [2004] asignan «pesos» a cada adjetivo en función de la posición que ocupa en el documento. Para Taboada, «no todos los adjetivos se crean de la misma forma» [Voll y Taboada, 2007, p. 337], por lo tanto cada adjetivo presenta una «relevancia» diferente en función de la posición que ocupe en el comentario. Siguiendo esta misma idea Taboada, Brooke y Stede [2009] consideran que el tratamiento de la polaridad compete solamente a ciertos segmentos o zonas del texto las cuales se pueden determinar a partir del concepto de zona o «movimiento» [Swales, 1990]. Los investigadores describen una taxonomía de segmentos discursivos basada en el género textual de los comentarios sobre películas que utilizan en la detección de la polaridad de este tipo de documentos. La clasificación de las 27 zonas propuestas se realizó mediante varios algoritmos de aprendizaje automático: `Naive Bayes`, `Support Vector Machine` y `Linear Regression`. En este sentido, es importante indicar que el error más frecuente de este tipo de aproximación es incurrir en una excesiva segmentación de los textos puesto que, como las mismos Taboada, Brooke y Stede [2009] sostienen, un mayor número de segmentos comporta una reducción en el rendimiento de los clasificadores. Taboada, Brooke y Stede concluyen que la reducción en el peso de los párrafos descriptivos aumenta el rendimiento de los clasificadores. Este hallazgo confirma su idea de que no todos los segmentos son igualmente útiles para calcular la polaridad.

Farra y col. [2010] dividen los comentarios en pequeños fragmentos de texto (*chunks*) conformados por conjuntos de oraciones. Según los autores, las oraciones que expresan opinión suelen aparecer al principio o al final del comentario, mientras que las oraciones neutras dominan la parte media de estos textos. Además, según su análisis, las dos últimas oraciones contienen las conclusiones sobre el producto valorado, en su caso, películas. Su trabajo, que es para el idioma árabe, se apoya en un análisis sintáctico superficial de los textos.

A continuación, presento las conclusiones generales que se obtienen de esta revisión de los antecedentes y del estado del arte.

2.4 CONCLUSIONES

En este capítulo he revisado varias aproximaciones al tratamiento de la polaridad en textos de opinión. Para ello, he tomado como eje de la descripción los diferentes métodos que se usan para representar los comentarios como un vector de características. Cada uno de estos métodos tiene sus ventajas y sus limitaciones.

La bolsa de palabras es un modelo de fácil implementación y rendimiento razonable pero presenta una alta dimensionalidad, un elevado nivel de dispersión e incapacidad de captar correctamente el contexto. Aunque este modelo puede optimizarse para buscar similitudes entre palabras (Análisis Semántico Latente) sigue sin tener en cuenta el orden o el rol de las mismas. Los *n-gramas* tienen en cuenta el orden de las palabras pero no su semántica. Los vectores sintácticos permiten establecer las relaciones estructurales que se establecen de manera discontinua entre diferentes elementos de la oración pero su utilidad para representar la polaridad ha sido cuestionada y, además, son complejos de generar. Las aproximaciones basadas en la estructura del discurso permiten una interpretación global de los textos pero suelen basarse en *parsers* discursivos complejos y poco fiables.

Considero que de estas aproximaciones, las que se apoyan en información discursiva pueden aportar soluciones alternativas e innovadoras al tratamiento de la polaridad de los comentarios sobre productos. Dentro de estas aproximaciones, las que se basan en la segmentación de los textos (*discourse zoning*) son las más viables puesto que no presentan las limitaciones de los *parsers* discursivos. Por lo tanto, mi propuesta irá orientada a la segmentación automática de los comentarios sobre productos con el propósito de identificar la función que cada tipo de segmento desempeña en la asignación de la polaridad. Para la segmentación de los comentarios me baso en su género discursivo [12] por considerar que el género determina la estructura discursiva asociada a un determinado propósito comunicativo.

La diferencia principal entre mi aproximación y las precedentes es que introduce una fase previa al análisis de la polaridad centrada en el tratamiento de la estructura discursiva de los comentarios. Este tratamiento no lo realizo a partir de complejos *parsers* discursivos, sino mediante el análisis lingüístico de los textos de opinión. Considero que el conocimiento sobre el funcionamiento lingüístico de los comentarios es indispensable para describir su estructura discursiva. Para acceder a dicho conocimiento se deben anotar los textos de opinión con información lingüística básica. En el capítulo siguiente describo los recursos y herramientas que he empleado para la anotación lingüística de los comentarios sobre productos.

12 Según diversos autores, el género discursivo es un factor relevante para la detección del afecto y las emociones que requiere un estudio más específico: [H. Li y col., 2012], [Pajupuu y col., 2012], [Ganu y col., 2010], [Goldberg y col., 2009], [Finn y Kushmerick, 2003, 2006], [Wiebe y col., 2004] y [Y.-B. Lee y Myaeng, 2002].

3 RECURSOS Y HERRAMIENTAS

Resumen

En este capítulo describo las herramientas y los recursos de los que obtengo la información lingüística necesaria para detectar la estructura discursiva de los comentarios sobre productos. Puesto que el objetivo final del tratamiento lingüístico de los comentarios es el cálculo de su polaridad, la descripción de estas herramientas y recursos la realizo siguiendo las tres etapas que caracterizan cualquier estudio experimental en el marco del Análisis de los Sentimientos y la Minería de Opiniones: la compilación, el procesamiento y la clasificación de los textos de opinión.

3.1 INTRODUCCIÓN

El objetivo de este capítulo es presentar los diferentes recursos y herramientas que he empleado para el cálculo de la polaridad de los comentarios mediante el análisis de su estructura discursiva. Esta presentación se realiza siguiendo el orden establecido por las tres etapas que, de acuerdo con Rahate y M [2013], caracterizan cualquier estudio experimental enfocado al análisis de sentimientos y que hago extensivo al análisis de la polaridad de los comentarios sobre productos: la compilación y preparación del corpus de comentarios, el procesamiento lingüístico de los comentarios y la clasificación de la polaridad de los comentarios. Estas tres etapas quedan representadas en la Figura 2 conjuntamente con los recursos y herramientas que he aplicado para su ejecución. En los siguientes apartados de este capítulo definiré cada una de estas tres etapas a partir de la descripción de los recursos y herramientas que involucra su ejecución.

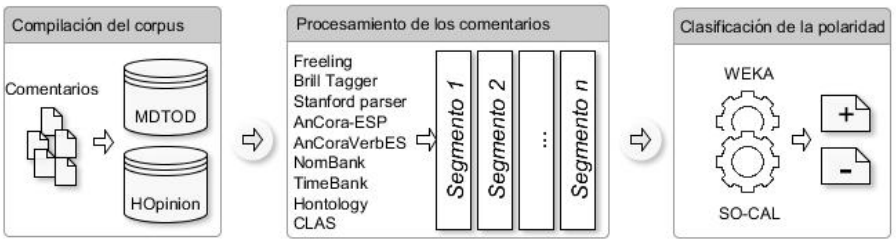


Figura 2: Recursos y herramientas utilizados en cada una de las etapas del tratamiento computacional de la polaridad.

3.2 COMPILACIÓN DE LOS CORPUS DE COMENTARIOS

La primera etapa que ejecuto como parte del tratamiento computacional de la polaridad en comentarios sobre productos es la compilación y preparación del corpus de textos valorativos o de opinión (ver Figura 2). El objetivo de esta etapa es obtener un corpus de comentarios sobre productos que sea representativo y de calidad. Por representatividad se entiende que el conjunto de documentos seleccionado proporcione un modelo verosímil de las propiedades lingüísticas de una población entera. La calidad del corpus tiene que ver con la preparación y adecuación de los textos para facilitar su análisis. Los experimentos de esta tesis se han realizado sobre dos corpus diferentes: un corpus, en español, compilado exclusivamente para esta investigación y un corpus, en inglés, compilado por Cruz Mata [2012] como parte de su tesis doctoral. A continuación describo ambos recursos.

3.2.1 HOpinion

HOpinion es un corpus de comentarios en castellano sobre hoteles que ha sido anotado con la polaridad de los textos. HOpinion contiene 18.490 comentarios (2.388.848 palabras) provenientes de las webs de TripAdvisor[1] y Booking[2]. El corpus[3] contiene comentarios fechados entre los años 2004 al 2011, donde los usuarios han puntuado los hoteles en un rango del 1 al 5, siendo 1 la peor puntuación y 5 la mejor. Algunos de los comentarios del corpus HOpinion van acompañados de información lingüística y todos, de metadatos.

En cuanto a la información lingüística, 4.740 comentarios están anotados con información morfosintáctica (Part Of Speech, POS)[4] que ha sido revisada manualmente por un grupo de lingüistas. Estos comentarios están codificados en formato CSV (ver Cuadro 2) y también en formato XML para facilitar su procesamiento y anotación (ver Cuadro 3).

	Forma	Lema	PartOfSpeech
2	Gracias_a	gracias_a	SPS00
3	dios	dios	NCMS000
4	solo	s{\'o}lo	RG
5	pasamos	pasar	VMIS1P0
6	una	uno	DI0FS0
7	noche	noche	NCFS000
8	en	en	SPS00
9	este	este	DD0MS0
10	hotel	hotel	NCMS000
11	...		

Cuadro 2: Fragmento de un comentario del corpus HOpinion anotado con información morfosintáctica y en formato CSV.

Por su parte, los metadatos aportan información sobre el comentario, el usuario y el ítem (hotel). Todo comentario contiene una serie de metadatos que se usan para su identificación: el identificador (ID) relativo del comentario, el ID absoluto del comentario, el alias del usuario, el ID relativo del

1 http://www.tripadvisor.es/
2 http://www.booking.com/
3 Recopilado en los años 2010 y 2011.
4 Las etiquetas POS están expresadas según el estándar EAGLES para la anotación morfosintáctica de lexicones y corpus en lenguas europeas.

```
1  <?xml version="1.0" encoding="UTF-8" standalone="no"?>
2  <!DOCTYPE article SYSTEM "hopinion.dtd">
3  <article comment="Imported into a project" morpho="automatic:02/07/12
       11:23" source="C:\Users\Usuari\Desktop\150tagged\07017
       _morfologic_8084448_saezb.txt">
4    <sentence>
5      <s complex="no" lem="gracias_a" name="s" pos="spsoo" postype="
           preposition" wd="Gracias_a"/>
6      <n gen="m" lem="dios" name="n" num="s" pos="ncmsooo" postype="
           common" wd="dios"/>
7      <r lem="s{\'o}lo" name="r" pos="rg" wd="solo"/>
8      <v lem="pasar" mood="indicative" name="v" num="p" person="1" pos="
           vmis1po" postype="main" tense="past" wd="pasamos"/>
9      <d gen="f" lem="uno" name="d" num="s" pos="diofso" postype="
           indefinite" wd="una"/>
10     <n gen="f" lem="noche" name="n" num="s" pos="ncfsooo" postype="
           common" wd="noche"/>
11     <s complex="no" lem="en" name="s" pos="spsoo" postype="preposition
           " wd="en"/>
12     <d gen="m" lem="este" name="d" num="s" pos="ddomso" postype="
           demonstrative" wd="este"/>
13     <n gen="m" lem="hotel" domain="yes" name="n" num="s" pos="ncmsooo"
           postype="common" wd="hotel"/>
14     ...
```

Cuadro 3: Fragmento de un comentario con información morfosintáctica en formato XML del corpus HOpinion.

hotel que está siendo valorado, la puntuación que da el usuario a dicho hotel, la fecha de publicación del comentario y, finalmente, el texto original del comentario. La nomenclatura usada para los IDs es la siguiente: r* (identificador del comentario o *review*), g* (localización geográfica del hotel) y d* (identificador del hotel). A continuación se muestra un ejemplo de la manera en que se ha codificado esta información:

> r8084448 | g187448-d232751-r8084448 | saezb | d232751 | 1 | 9 jul 2007 | Gracias a dios solo pasamos una noche en este hotel, las habitaciones viejisimas!! duchas oxidadas, albornoces rotos, sabanas de la cama quemadas, la ventana de la habitacion no se podia abrir (estaba rota) el mando de la television tenia los numeros borrados....Eso si seleccionamos una habitacion superor (ellos la llaman CLUB) porque teniamos cama de 2*2 metros (realmente era una de 1,90 * 1,70) y por ser mas grande y con mejor decoracion.... No se como seria la de las demas habitaciones. Lamentable!!, no se lo recomiento a nadie. Viejo, malo, sucio y la atencion patetica.

En el proceso de compilación, también se ha recuperado la información relativa a cada usuario: alias, género, franja de edad, ciudad de residencia y una pequeña descripción del tipo de viajes que le agrada realizar. El siguiente es una entrada del documento CSV que contiene esta información:

> Acetuchin | Mujer | 25-34 | Madrid | Mi estilo de viaje: Dándose un lujo de vez en cuando.

Finalmente, los metadatos con la descripción de cada uno de los hoteles que han sido valorados son los siguientes: el ID que vincula al hotel con una localización geográfica específica, el ID relativo del hotel, el nombre del hotel, el tipo de alojamiento y el número de estrellas. Además, se ha recuperado las veces que los viajeros han puntuado el hotel como (excelente), (muy bueno), (normal), (malo) y (pésimo), y el tipo de viajes que se han hecho a este hotel: (de negocios), (en pareja), (en familia), (con amigos) y (en solitario).

Finalmente, se recuperó la descripción de los servicios con los que cuenta el hotel:

> g1006045 | d1144505 | Hotel Parador de Cruz de Tejeda | complejo turístico | 4 | (32)(26)(3)(2)(0) | (4)(51)(1)(0)(2) | Restaurante, Bar, Recepción 24 horas, Prensa, Terraza, Habitaciones no fumadores...

Para mayor claridad, la información relacionada con la localización geográfica de los hoteles (ID) se ha detallado de forma independiente:

> g1006045 | [Europa][España][Islas Canarias][Gran Canaria][Tejeda]

HOpinion se puede descargar gratuitamente de la web del Centro de Lenguaje y Computación[5] (CLiC) en forma de texto plano (CSV) y base de datos. Los archivos que se obtiene tras la descarga son:

1. MYSQLFILES: archivos **.sql** para importar a la base de datos de MySQL y un archivo **.mwb** para MySQL Workbench[6] con la estructura de la base de datos.
2. TABLES_DB: los archivos originales en texto plano (.txt).
3. MORPHO: 4.750 archivos analizados morfológicamente.
4. TXT-FULL: zip con los comentarios en texto plano.

3.2.2 Multi-Domain Taxonomy-based Opinion Dataset

Multi-Domain Taxonomy-based Opinion Dataset[7] (MDTOD) [Cruz Mata, 2012] es un corpus de comentarios en inglés sobre coches, hoteles y auriculares que ha sido anotado con la polaridad de los textos y de las características opinables de los objetos que valoran los usuarios. El corpus contiene 587 comentarios sobre auriculares, 972 comentarios sobre coches y 988 comentarios sobre hoteles. En total contiene 2.547 comentarios extraídos de la web de Ciao.com[8]. Al igual que HOpinion, los comentarios están puntuados por los usuarios en un rango del 1 al 5, siendo 1 la peor puntuación y 5 la mejor. El corpus se compone de los comentarios anotados y una taxonomía con las características del dominio.

Un ejemplo con el tipo de anotación utilizada en el corpus MDTOD se puede ver en el Cuadro 4. La anotación incluye, en la línea 2, la puntuación dada por el usuario al objeto (auricular, coche u hotel); en la línea 4, el título del comentario; y en las líneas 7 y 10, los *pros* y los *contras*. Además, el texto del comentario (línea 13 en adelante) está segmentado por oraciones y cada oración en opiniones. En la anotación que se hace de las opiniones (líneas 17, 18 y 23) se considera, la polaridad (*polarity*) positiva (+) y negativa (–), la característica (*feature*), la referencia al número de la palabra que dentro de la oración expresa la característica (*featWords*) y la polaridad (*opWords*).

El corpus MDTOD incluye una relación de los conceptos de cada dominio sobre los cuales es posible opinar («características opinables del objeto») como pueden ser «decoración» (*decoration*), «suelo» (*ground*) o escaleras (*stairway*) en el caso de los hoteles. Al ser una relación ordenada, jerarquizada y codificada de conceptos, adquiere la condición de taxonomía. En el Cuadro 5 tenemos un fragmento de la taxonomía sobre hoteles en donde, por

5 http://clic.ub.edu/
6 http://www.mysql.com/products/workbench/
7 http://www.lsi.us.es/~fermin/index.php/Datasets
8 http://www.ciao.com/

ejemplo, el concepto «*building*» se compone de los conceptos subordinados «*decoration*», «*grounds*», «*lobby*», «*patio*», «*roof*» y «*stairways*».

```
<?xml version="1.0" encoding="UTF-8"?>
<review id="983" item="Lakeside Inn" rating="4">
  <title>
    <sentence id="1"> Rio(1) legend(2) </sentence>
  </title>
  <pros>
    <sentence id="1"> Beautiful(1) interior(2) </sentence>
  </pros>
  <cons>
    <sentence id="1"> Faded(1) exterior(2) </sentence>
  </cons>
  ...
  <text>
  ...
    <sentence id="2">
      The(1) hotel(2) exterior(3) is(4) beautiful(5) ,(6) it(7) 's(8) the(9) best-
          looking(10) hotel(11) in(12) Rio(13) .(14)
      <opinion polarity="+" feature="hotel" featWords="2" opWords="5"/>
      <opinion polarity="+" feature="hotel" featWords="11" opWords="10"/>
    </sentence>
    ...
    <sentence id="7">
      Bathroom(1) was(2) gloomy(3) and(4) fixtures(5) were(6) outdated(7) .(8)
      <opinion polarity="-" feature="bathroom" featWords="1" opWords="3"/>
    </sentence>
    ...
```

Cuadro 4: Fragmento de un comentario anotado con la polaridad de las opiniones sobre las características en formato XML del corpus MDTOD.

```
<?xml version="1.0" encoding="UTF-8"?>
<taxonomy productClass="hotel">
  <featureRoot name="hotel" featWords="complex,hotel,hotels,motel,motels,resort,
      resorts">
    <feature name="building" featWords="architecture,building,buildings">
      <feature name="decoration" featWords="decor,decoration,decorations,exterior,
          landscaping,scenery,setting,site,surroundings"/>
      <feature name="grounds" featWords="floor,floors,grounds"/>
      <feature name="lobby" featWords="entrance,hall,halls,hallway,hallways,lobbies,
          lobby"/>
      <feature name="patio" featWords="patio,patios"/>
      <feature name="roof" featWords="ceilings,roof"/>
      <feature name="stairways" featWords="stairs,stairways"/>
    </feature>
      ...
```

Cuadro 5: Fragmento de la taxonomía sobre hoteles del corpus MDTOD de Cruz Mata [2012].

3.3 RECURSOS Y HERRAMIENTAS PARA EL PROCESAMIENTO LINGÜÍSTICO DE LOS COMENTARIOS

La segunda etapa que ejecuto como parte del tratamiento computacional de la polaridad en comentarios sobre productos es el procesamiento lingüístico de los textos de opinión (ver Figura 2 al principio del capítulo). El propósito de esta etapa es enriquecer los comentarios con información léxica,

morfológica y sintáctica que pueda utilizarse para identificar y tratar los principales segmentos que los caracterizan.

Las herramientas utilizadas para el análisis lingüístico de los comentarios son los léxicos (diccionarios), los analizadores morfológicos y los *parsers* sintácticos. El objetivo de los recursos léxicos es enriquecer la descripción de los comentarios con información específica como puede ser la orientación semántica de las palabras o la detección de los términos y expresiones propias de un dominio. El objetivo de los analizadores morfológicos es la identificación de las palabras y su categorización morfológica. El objetivo de los *parsers* sintácticos es establecer las relaciones de dependencia entre palabras a nivel de frase.

3.3.1 Herramientas

En los apartados siguientes describo las herramientas empleadas en la tesis para el análisis morfológico y sintáctico.

3.3.1.1 *Freeling*

Freeling[9] [Padró, 2011] es una librería de código abierto destinada al análisis y la anotación lingüística de textos que empleo para enriquecer con **información morfosintáctica** los comentarios provenientes del corpus HOpinion.

Freeling proporciona una amplia gama de servicios de análisis lingüístico para diversos idiomas, entre ellos el castellano y el inglés. Adicionalmente, está orientado a facilitar la integración con aplicaciones de niveles superiores y al desarrollo de aplicaciones para el Procesamiento del Lenguaje Natural, lo que lo hace fácilmente configurable. En este sentido, es posible codificar los textos directamente en formato XML mediante el *plugin* de AnCoraPipe [Bertran y col., 2010]. Por último, está fuertemente orientado a aplicaciones del mundo real por lo que es un recurso veloz y robusto.

En el Cuadro 6 presento un ejemplo de la salida del etiquetador morfológico de *Freeling*, aplicado un comentario sobre hoteles del corpus HOpinion.

3.3.1.2 *Brill Tagger*

Brill Tagger[10] [Brill, 1994] es un sistema de **etiquetado morfológico** de textos en inglés, que he utilizado para el análisis de los comentarios provenientes del corpus MDTOD.

Brill Tagger es un sistema de etiquetado basado en reglas de transformación que han sido inferidas a partir de un corpus de entrenamiento. El etiquetador alcanza un nivel de rendimiento comparable al de los analizadores estocásticos mediante la codificación de un pequeño y simple conjunto de reglas que obtienen la información lingüística de manera directa y eficiente. Las reglas fueron generadas mediante un proceso de entrenamiento a partir del corpus ITU[11]. Este corpus contiene el manual de la *International Telecommunications Union CCITT*, conocido como *The Blue Book*, en su versión en español e inglés. *Brill Tagger* consta de tres módulos: un etiquetador léxico, un etiquetador de palabras desconocidas y un etiquetador contextual.

Debido al formato que *Brill Tagger* utiliza para etiquetar los textos, los comentarios procesados con esta herramienta se emplearon como entrada al

9 http://nlp.lsi.upc.edu/freeling/
10 http://cst.dk/online/pos_tagger/uk/
11 http://www.itu.ch/itudoc/itu-t/rec/g/g700-799.html

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!DOCTYPE article SYSTEM "hopinion.dtd">
<article comment="Imported into a project" morpho="automatic:02/07/12
    11:23" source="C:\Users\Usuari\Desktop\150tagged\01934
    _morfologic_20303931_Nelly021.txt">
  <sentence>
    <v lem="estar" mood="indicative" name="v" num="s" person="1" pos="
        vmis1so" postype="main" tense="past" wd="Estuve"/>
    <d gen="f" lem="uno" name="d" num="s" pos="diofso" postype="
        indefinite" wd="una"/>
    <n gen="f" lem="noche" name="n" num="s" pos="ncfs000" postype="
        common" wd="noche"/>
    <s complex="no" lem="en" name="s" pos="spsoo" postype="preposition
        " wd="en"/>
    <d gen="m" lem="el" name="d" num="s" pos="daomso" postype="article
        " wd="el"/>
    <n gen="m" lem="hotel" name="n" num="s" pos="ncms000" postype="
        common" wd="hotel"/>
    <c lem="y" name="c" pos="cc" postype="coordinating" wd="y"/>
    <r lem="no" name="r" pos="rn" postype="negative" wd="no"/>
    <p gen="c" lem="yo" name="p" num="s" person="1" pos="pp1cs000"
        postype="personal" wd="me"/>
    <v lem="parecer" mood="indicative" name="v" num="s" person="3" pos
        ="vmis3so" postype="main" tense="past" wd="parecio"/>
    <d gen="m" lem="uno" name="d" num="s" pos="diomso" postype="
        indefinite" wd="un"/>
    <z lem="4" name="z" pos="z" wd="4"/>
    <n gen="c" lem="estrella" name="n" num="p" pos="nccp000" postype="
        common" wd="estrellas"/>
    ...
```

Cuadro 6: Fragmento en formato XML de un texto del corpus HOpinion analizado con Freeling y codificado en XML mediante el *plugin* de AnCoraPipe.

Semantic Orientation CALculator System (SO-CAL), un software para la clasificación de textos por polaridad. En el Cuadro 7 se muestra un ejemplo de la salida del etiquetador *Brill Tagger* aplicado a un comentario del corpus MDTOD.

```
The/DT sound/JJ quality/NN is/VBZ not/RB bad/JJ ,/, but/CC the/DT bass
    /NN sounds/VBZ a/DT little/JJ muffled/NN ./.
```

Cuadro 7: Ejemplo del análisis generado por Brill Tagger a partir de la oración «*The sound quality is not bad, but the bass sounds a little muffled*», extraída del corpus MDTOD.

3.3.1.3 *Parser de Stanford*

El Parser de la Universidad de Stanford (*Stanford Parser*) [Klein y Manning, 2003] es un *parser* de dependencias que he empleado para enriquecer con **información sintáctica** los comentarios sobre productos del corpus MDTOD. Una dependencia sintáctica es una relación jerárquica que se establece entre pares palabras, una de las cuales desempeña la función de núcleo y la otra de modificador.

El Parser de Stanford está implementado en Java y su funcionamiento se basa en un enfoque probabilístico. La ventaja principal de los analizadores basados en dependencias es que suelen ser más rápidos que los analizadores basados en constituyentes. En cuanto a su rendimiento, el Parser de Stanford ha demostrado ser una herramienta robusta. Comelles y col. [2010] efectúan una evaluación cuantitativa y cualitativa de un grupo de analizadores de constituyentes y de dependencias en la que el Parser de Stanford presenta el

mejor rendimiento. En el Cuadro 8 muestra un ejemplo de las dependencias identificadas por el Parser de Stanford a partir de una oración extraída del corpus MDTOD.

```
1  det(quality-3, The-1)
2  amod(quality-3, sound-2)
3  nsubj(bad-6, quality-3)
4  cop(bad-6, is-4)
5  neg(bad-6, not-5)
6  root(ROOT-0, bad-6)
7  cc(bad-6, but-8)
8  det(bass-10, the-9)
9  nsubj(sounds-11, bass-10)
10 conj(bad-6, sounds-11)
11 det(muffled-14, a-12)
12 amod(muffled-14, little-13)
13 dobj(sounds-11, muffled-14)
```

Cuadro 8: Ejemplo de las dependencias generadas por el Parser de Stanford a partir de la oración «*The sound quality is not bad, but the bass sounds a little muffled*», extraída del corpus MDTOD.

3.3.1.4 *Computerized Linguistic Analysis System*

Computerized Linguistic Analysis System[12] (CLAS) [Pakhomov y col., 2010] es una aplicación que mide la complejidad sintáctica de los textos y que empleo para enriquecer el corpus MDTOD con **información sobre la complejidad sintáctica** de sus oraciones.

CLAS ha sido diseñada en lenguaje Java y se compone de cuatro módulos que se ejecutan de manera secuencial. El primer módulo es el Segmentador (*Tokenizer*), encargado de identificar los *tokens* individuales del texto (palabras, signos de puntuación, fechas, etc.). El segundo módulo es el detector de oraciones que se encarga de detectar los límites de las oraciones. El tercer módulo realiza el análisis sintáctico de cada oración mediante una llamada al Parser de Stanford. Finalmente, el cuarto módulo es el *Complexity Score Generator* que calcula varias métricas asociadas con la complejidad sintáctica.

He seleccionado CLAS dentro de los diferentes sistemas que existen para calcular la complejidad sintáctica puesto que su rendimiento ha sido evaluado de forma positiva en múltiples trabajos[13]. Además, CLAS ha sido rediseñado (junio de 2011) para mejorar su rendimiento y facilitar su uso. Por último, aunque no menos importante, CLAS es el sistema que incluye un número mayor de métricas lo cual resulta útil de cara a enriquecer el vector de características con el que se representa cada uno de los comentarios.

A manera de ejemplo, en el Cuadro 9 se puede ver la salida del *Computerized Linguistic Analysis System* tras analizar una oración extraída del corpus MDTOD. Los valores que aparecen en la línea 2 del ejemplo corresponden a diferentes métricas de la complejidad sintáctica que se detallan en la Sección 5.5.

```
1 It has ample trunk space and looks to die for.
2 |2|1.0909091|12.0|2.4545455|27.0|2.0|16|2|1|0|3|0|1|1|0
```

Cuadro 9: Análisis de la complejidad sintáctica de la oración «*It has ample trunk space and looks to die for*» mediante CLAS.

12 http://rxinformatics.umn.edu/clas.html

13 Ver la lista de publicaciones en la página del proyecto: http://rxinformatics.umn.edu/publications.html

3.3.2 Recursos

En los apartados siguientes describo los recursos que he utilizado para enriquecer la anotación de los comentarios con información específica y del dominio.

3.3.2.1 *AnCora-ESP*

AnCora-ESP[14] [Taulé y col., 2008] es un corpus de 500.000 palabras constituido principalmente por artículos periodísticos en español que, al presentar unas características lingüísticas diferentes al de los comentarios sobre productos, he empleado para contrastar los datos extraídos de los corpus MDTOD y HOpinion.

AnCora-ESP está anotado con **información morfológica** (lema y categoría), sintáctica (constituyentes y funciones), semántica (estructura argumental, roles semánticos, entidades nombradas y sentidos nominales de WordNet) y pragmática (correferencia).

Un fragmento de uno de los textos del corpus AnCoraESP aparece en el Cuadro 10. Como se puede observar, la codificación para HOpinion (Cuadro 3), MDTOD (Cuadro 4) y AnCora-ESP (Cuadro 10) es fácilmente equiparable, lo que ha facilitado la explotación de estos recursos.

```
1  <?xml version="1.0" encoding="UTF-8" standalone="no"?>
2  <article>
3  <sentence>
4   <d gen="f" lem="mucho" num="p" pos="diofpo" postype="indefinite" wd="
       Muchas"/>
5   <n gen="f" lem="vez" num="p" pos="ncfp000" postype="common" sense
       ="16:05449233" wd="veces"/>
6   <p gen="c" lem="{\'e}l" num="c" person="3" pos="p0300000" wd="se"/>
7   <v els="a2" lem="tirar" mood="indicative" num="s" person="3" pos="
       vmii3s0" postype="main" tense="imperfect" wd="tiraba"/>
8   <s contracted="yes" gen="m" lem="al" num="s" pos="spcms" postype="
       preposition" wd="al"/>
9   <n gen="m" lem="suelo" num="s" pos="ncms000" postype="common" sense
       ="16:02699889" wd="suelo"/>
10  <s lem="para" pos="sps00" postype="preposition" wd="para"/>
11  <v els="a2" lem="rodear" mood="infinitive" pos="vmn0000" postype="
       main" wd="rodear"/>
12  <s lem="con" pos="sps00" postype="preposition" wd="con"/>
13  <d gen="m" lem="el" num="p" pos="da0mp0" postype="article" wd="los"/>
14  <n gen="m" lem="brazo" num="p" pos="ncmp000" postype="common" sense
       ="16:04310435" wd="brazos"/>
15  <d gen="m" lem="el" num="p" pos="da0mp0" postype="article" wd="los"/>
16  <n gen="m" lem="tobillo" num="p" pos="ncmp000" postype="common" sense
       ="16:04322434" wd="tobillos"/>
17  <s lem="de" pos="sps00" postype="preposition" wd="de"/>
18  <d gen="c" lem="su" num="s" person="3" pos="dp3cs0" postype="
       possessive" wd="su"/>
19  <n gen="f" lem="madre" num="s" pos="ncfs000" postype="common" sense
       ="16:07419862" wd="madre"/>
20     ...
```

Cuadro 10: Fragmento en formato XML de un texto periodístico del corpus AnCora-ESP.

14 http://clic.ub.edu/corpus/ancora

3.3.2.2 *AnCoraVerbES*

AnCoraVerbES[15] (versión 2.0.3) [Aparicio y col., 2008] es un léxico verbal que empleo para enriquecer los comentarios provenientes del HOpinion con **información sobre las clases eventivas** de los verbos.

AnCoraVerbES contiene los esquemas sintáctico-semánticos asociados a los diferentes sentidos de 2.248 verbos en español (3.938 sentidos). Cada uno de estos esquemas tiene asociada la clase semántica, la estructura argumental, con papeles temáticos asignados a cada argumento, y su proyección a la estructura sintáctica. Los esquemas sintáctico-semánticos se distinguen de acuerdo con las cuatro clases eventivas o *Aktionsart*: realización (A), logro (B), estado (C) y actividad (D).

En el Cuadro 11 se muestra la información relacionada con el verbo «reforzar»: en la línea 2 aparecen el lema («reforzar») y la categoría («*verb*»), los diferentes sentidos asociados a las clases semánticas correspondientes aparecen en las líneas 4 y 8. En el ejemplo, el predicado «reforzar» se puede clasificar como una acción **A** (ej. «la operación refuerza su liderazgo») o un logro **B** (ej. «si dos neuronas se activan, sus conexiones se refuerzan»).

```
1  <?xml version="1.0" encoding="UTF-8"?>
2    <lexentry lemma="reforzar" lng="es" type="verb">
3      <sense id="1">
4        <frame default="yes" lss="A11.transitive-causative" type="default">
5          <argument argument="argo" function="suj" thematicrole="cau"/>
6          <argument argument="arg1" function="cd" thematicrole="tem"/>
7          ...
8        <frame lss="B21.unaccusative-state" type="anticausative">
9          <argument argument="arg1" function="suj" thematicrole="tem"/>
10         <argument argument="argM" function="cc" thematicrole="adv"/>
11         ...
```

Cuadro 11: Fragmento de la entrada léxica para el verbo «reforzar» en AnCoraVerbES.

3.3.2.3 *NomBank*

NomBank[16] [Meyers y col., 2004] es un corpus etiquetado con los roles semánticos a nivel nominal que empleo para enriquecer el corpus MDTOD con la anotación de las **nominalizaciones deverbales**, es decir, nombres que mantienen una relación morfológica y/o semántica con verbos de los cuales heredan su estructura argumental: $\texttt{abducir}_v \rightarrow \texttt{abducción}_n$, $\texttt{amenazar}_v \rightarrow \texttt{amenaza}_n$, $\texttt{embalar}_v \rightarrow \texttt{embalaje}_n$.

NomBank incluye la anotación semántica de la estructura argumental de los nombres deverbales o nominalizaciones deverbales que aparecen en el corpus PennTreeBank (5.000 aproximadamente) [Palmer y col., 2005]. NomBank comparte el esquema de anotación del proyecto PropBank, en el que se realizó la anotación de la estructura argumental de los predicados verbales del PennTreeBank. En ambos corpus se utilizan las etiquetas *argo*, *arg1*, *arg2*, *arg3* y *arg4* para expresar, mediante orden numérico incremental, el grado de proximidad de un argumento con respecto a su predicado. Los adjuntos, por su parte, se etiquetan como *argM*.

Un ejemplo de entrada léxica, también llamada *frame*, del NomBank aparece en el Cuadro 12.

15 http://clic.ub.edu/corpus/ancoraverb_es
16 http://nlp.cs.nyu.edu/meyers/NomBank.html

```
1  <!DOCTYPE frameset SYSTEM "frameset.dtd">
2  <frameset>
3    <predicate lemma="abduction">
4      <roleset id="abduction.01" name="to carry someone off by force" source="verb–
           abduct.01">
5        <roles>
6          <role descr="agent" n="0"></role>
7          <role descr="person kidnapped" n="1"></role>
8        </roles>
9        <example name="autogen1">
10         <text>her alleged abduction</text>
11         <arg n="1">her</arg>
12         <arg f="ADV" n="M">alleged</arg>
13         <rel>abduction</rel>
14       </example>
15     </roleset>
16   </predicate>
17 </frameset>
```

Cuadro 12: Un ejemplo de entrada léxica (*frame*) en NomBank.

3.3.2.4 *TimeBank*

TimeBank[17] (versión 1.2) [Pustejovsky y col., 2003] es un corpus del inglés compuesto por 183 documentos anotados con **información temporal y eventiva** (7.935 eventos) que he aplicado para identificar las expresiones temporales presentes en los comentarios del corpus MDTOD.

TimeBank utiliza dos estructuras de datos para especificar, mediante el lenguaje de especificación TimeML, las principales expresiones temporales del inglés: TIMEX3 y SIGNAL.

1. TIMEX3: se usa para anotar expresiones temporales: «*1 January of the following year*», «yesterday», «at 7:30 in the morning», «the next year».
2. SIGNAL se usa para anotar marcadores de relaciones temporales: «*before*», «*after*», «*during*».

En el Cuadro 13 aparece un fragmento de un texto de TimeBank anotado mediante el lenguaje TimeML. Este fragmento corresponde a la oración «*The warrants may be exercised until 90 days after their issue date*». Como se puede observar, en la línea 1 se indica la fecha de publicación del documento que servirá como referencia temporal para «normalizar» los eventos descritos en el texto; en las líneas 5 y 7 se caracterizan dichos eventos («*exercised*» y «*issue*»); y en la línea 9 se especifica que el evento «*exercised*» es posterior («AFTER») al evento «*issue*».

3.3.2.5 *Hontology*

Hontology [Silveira y col., 2012] es una ontología[18] multilingüe para el sector de hoteles y alojamientos que he empleado para identificar los **términos de dominio** presentes en el corpus HOpinion.

Hontology ha sido definida mediante el Lenguaje de Ontologías Web (*Web Ontology Language*, OWL). Los idiomas sobre los que se ha desarrollado la ontología son inglés, español, portugués y francés. Hontology contiene 282 conceptos organizados en 16 grandes categorías. La jerarquía de conceptos tiene una profundidad máxima de cinco niveles. El Cuadro 14 se presenta un fragmento de Hontology: en la línea 4 se está declarando un concepto, en

17 `http://www.timeml.org/site/timebank/timebank.html`

18 Una ontología es un conjunto de términos y relaciones entre dichos términos que se usan para representar de manera exhaustiva y rigurosa un dominio dado [Staab y Studer, 2009].

las líneas 8 y 9 se presentan dos conceptos relacionados mediante el axioma `SubClassOf`, en 13 y 14 aparecen dos clases disjuntas, y en 18 y 19 se presenta un axioma del tipo `SubDataPropertyOf`. Estos ejemplos ayudan a entender como están dispuestos los conceptos del dominio en la ontología.

```
1  <TIMEX3 tid="t30" type="DATE" value="1989-10-26"
       temporalFunction="false" functionInDocument="
       CREATION_TIME">10/26/89</TIMEX3>
2  ...
3  The warrants may be <EVENT eid="e14" class="OCCURRENCE" stem
       ="exercise">exercised</EVENT> until <TIMEX3 tid="t31"
       type="DURATION" value="P90D" temporalFunction="false"
       functionInDocument="NONE">90 days</TIMEX3> <SIGNAL sid="
       s16">after</SIGNAL> their <EVENT eid="e32" class="
       OCCURRENCE" stem="issue">issue</EVENT> date.
4  ...
5  <MAKEINSTANCE eventID="e14" eiid="ei88" tense="NONE" aspect
       ="PERFECTIVE" polarity="POS" pos="VERB" modality="may"/>
6  ...
7  <MAKEINSTANCE eventID="e32" eiid="ei89" tense="NONE" aspect
       ="NONE" polarity="POS" pos="NOUN"/>
8  ...
9  <TLINK lid="l4" relType="AFTER" eventInstanceID="ei88"
       relatedToEventInstance="ei89"/>
10 ...
```

Cuadro 13: Fragmento de un texto de TimeBank anotado mediante el lenguaje TimeML.

```
1  <?xml version="1.0"?>
2  ...
3   <Declaration>
4    <Class IRI="http://www.semanticweb.org/ontologies/2010/3/
         hotel-15072010.owl#Accommodation"/>
5   </Declaration>
6   ...
7   <SubClassOf>
8    <Class IRI="http://www.semanticweb.org/ontologies/2010/3/
         hotel-15072010.owl#SuperiorLuxury"/>
9    <Class IRI="http://www.semanticweb.org/ontologies/2010/3/
         hotel-15072010.owl#HotelCategory"/>
10  </SubClassOf>
11  ...
12  <DisjointClasses>
13   <Class IRI="http://www.semanticweb.org/ontologies/2010/3/
         hotel-15072010.owl#Comfort"/>
14   <Class IRI="http://www.semanticweb.org/ontologies/2010/3/
         hotel-15072010.owl#Luxury"/>
15  </DisjointClasses>
16  ...
17  <SubDataPropertyOf>
18   <DataProperty IRI="http://www.semanticweb.org/ontologies
         /2010/3/hotel-15072010.owl#hasWiFi"/>
19   <DataProperty IRI="http://www.semanticweb.org/ontologies
         /2010/3/hotel-15072010.owl#hasInternetAccess"/>
20  </SubDataPropertyOf>
21  ...
```

Cuadro 14: Un fragmento del archivo OWL que contiene la ontología Hontology [Silveira y col., 2012].

3.4 HERRAMIENTAS PARA LA CLASIFICACIÓN DE LA POLARIDAD DE LOS COMENTARIOS

La tercera etapa que ejecuto como parte del tratamiento computacional de la polaridad en comentarios sobre productos es la clasificación de los textos de opinión (ver Figura 2). El objetivo de esta etapa es asignar los comentarios a una de las dos categorías posibles: positivos o negativos. Las herramientas que he empleado en esta tesis para la clasificación de los comentarios son el entorno Weka (*Waikato Environment for Knowledge Analysis*) [Witten y col., 1999] y el sistema SO-CAL (*Semantic Orientation CALculator System*) [Taboada, Brooke, Tofiloski y col., 2011]. A continuación describo ambas herramientas.

3.4.1 Waikato Environment for Knowledge Analysis

Waikato Environment for Knowledge Analysis[19] [Witten y col., 1999], más conocido por su acrónimo Weka, es un software de experimentación para la extracción de conocimiento. Weka soporta varias tareas típicas de minería de datos como el agrupamiento, la clasificación, la regresión o la selección de características. El software también incluye facilidades para la visualización de los datos. En esta tesis he aplicado Weka como herramienta para el aprendizaje supervisado, es decir, para construir modelos que permitan predecir la categoría de las instancias en función de una serie de características o datos de entrenamiento obtenidos en la fase de análisis.

Varias son las propiedades que hacen de Weka un software idóneo para este trabajo. En primer lugar, Weka pone a disposición de los usuarios, de una forma sencilla y transparente, los principales algoritmos de aprendizaje automático para su uso en tareas de minería de datos y clasificación automática. En segundo lugar, Weka es multiplataforma (ha sido desarrollado completamente en el lenguaje de programación Java) lo que únicamente obliga al usuario a disponer del intérprete de Java para poder ejecutar los diferentes algoritmos de aprendizaje. Finalmente, Weka se distribuye como software libre lo que ha permitido un crecimiento permanente gracias a los constantes aportes de la comunidad científica y es la herramienta de libre distribución preferida por los investigadores en Minería de Datos y Aprendizaje Automático. Weka se compone de una serie de algoritmos de aprendizaje automático, un conjunto de métodos de selección de atributos y permite aplicar la técnica de validación cruzada. A continuación describo estas características.

3.4.1.1 *Algoritmos de aprendizaje automático*

En esta tesis utilizo algunos de los algoritmos de clasificación supervisada presentes en Weka (versión 3.6). En general, un clasificador suministra una función que proyecta una instancia, especificada por una serie de características o atributos, en una serie de clases predefinidas [Sierra, 2006, p. 84]. La clasificación supervisada trabaja con datos etiquetados, es decir, cada uno de los casos o instancias de la muestra de entrenamiento tiene asociada una etiqueta o clase. Por lo tanto, cada instancia esta descrita por una serie de características y etiquetada con una clase. El objetivo principal de estos algoritmos es asignar una clase a casos desconocidos no etiquetados en base al aprendizaje realizado sobre una muestra de entrenamiento. En el Cuadro 15 presento los algoritmos de clasificación supervisada disponibles en el en-

19 http://www.cs.waikato.ac.nz/ml/weka/

torno Weka, algunos de los cuales se aplican en los experimentos de esta tesis.

Bayes

Son clasificadores estadísticos, que pueden predecir tanto las probabilidades del número de miembros de clase como la probabilidad de que una instancia dada pertenezca a una clase particular:
AODE, AODEsr, BayesNet, ComplementNaiveBayes, DMNBtext, HNB, BayesianLogisticRegression, NaiveBayes, NaiveBayesMultinomial, NaiveBayesMultinomialUpdateable, NaiveBayesMultinomialText, NaiveBayesSimple, NaiveBayesUpdateable, WAODE.

Funciones

Son algoritmos que maximizan o minimizan una función objetivo:
GaussianProcesses, IsotonicRegression, LeastMedSq, LibLINEAR, LibSVM, LinearRegression, Logistic, MLPClassifier, MLPRegressor, MultilayerPerceptron, PaceRegression, PLSClassifier, RBFNetwork, RBFClassifier, RBFRegressor, SimpleLinearRegression, SimpleLogistic, SGD, SGDText, SMO, SMOreg, SPegasos, SVMreg, VotedPerceptron, Winnow.

Perezosos

No generan modelos de manera explícita sino que almacenan las instancias de entrenamiento presentadas y no trabajan con ellas hasta el momento de realizar la clasificación de otras instancias:
IB1, IBk, KStar, LBR, LWL.

Árboles de clasificación

Un árbol de clasificación es un conjunto de reglas de decisión o condiciones organizadas en una estructura jerárquica:
ADTree, BFTree, DecisionStump, ExtraTree, FT, HoeffdingTree, Id3, J48, J48graft, LADTree, LMT, M5P, NBTree, RandomForest, RandomTree, REPTree, SimpleCart, UserClassifier, lmt.LogisticBase.

Reglas de decisión

Son clasificadores construidos a partir de métodos inductivos de reglas de tipo condicional:
ConjunctiveRule, DecisionTable, DTNB, FURIA, JRip, M5Rules, NNge, OneR, PART, Prism, Ridor, ZeroR.

Metaclasificadores

Son métodos que combinan las predicciones de varios modelos para clasificar nuevos ejemplos:
AdaBoostM1, AdditiveRegression, END, Bagging, ClassificationViaClustering, MetaCost, Stacking, ClassificationViaRegression, Dagging, CostSensitiveClassifier, Grading, GridSearch, CVParameterSelection, Decorate, AttributeSelectedClassifier, EnsembleSelection, FilteredClassifier, LogitBoost, MultiBoostAB, MultiClassClassifier, MultiClassClassifierUpdateable, MultiScheme, OneClassClassifier, OrdinalClassClassifier, RacedIncrementalLogitBoost, RandomCommittee, RandomSubSpace, RealAdaBoost, RegressionByDiscretization, RotationForest, StackingC, ThresholdSelector, Vote, ClassBalancedND, DataNearBalancedND, ND.

Misceláneos

HyperPipes, InputMappedClassifier, SerializedClassifier, MinMaxExtension, OLM, OSDL, VFI.

Cuadro 15: Algoritmos de clasificación del entorno Weka, algunos de los cuales he utilizado en esta tesis.

3.4.1.2 *Algoritmos de selección de atributos*

Considerando que el comportamiento de los clasificadores mejora cuando se eliminan los atributos no relevantes y redundantes, en el presente trabajo he aplicado varios métodos de selección de atributos disponibles en Weka. Los métodos de selección de atributos permiten elegir, del conjunto de características obtenidas en la etapa de análisis, aquellas que han sido más relevantes para clasificar los comentarios según su polaridad.

En este sentido, la función de la selección de atributos es triple: en primer lugar, permite mejorar el desempeño predictivo de los modelos de clasificación; en segundo lugar, favorece la construcción de modelos más eficientes mediante la ejecución de algoritmos más rápidos y menos costosos; y finalmente, mejora la comprensión de los modelos generados.

En Weka la selección de atributos se consigue mediante la aplicación de métodos evaluadores de subconjuntos de atributos (*Attribute Evaluators*) y de métodos de búsqueda (*Search Methods*). Los métodos evaluadores recorren un espacio de búsqueda de subconjuntos de atributos, evaluando subconjuntos completos de atributos. Los métodos de búsqueda, por su parte, ordenan los atributos de manera individual y eliminan los menos valorados. La lista completa de los métodos evaluadores y de búsqueda aparece en el Cuadro 16.

Métodos evaluadores	Métodos de búsqueda
ClassifierSubsetEval	BestFirst
CfsSubsetEval	ExhaustiveSearch
ChiSquaredAttributeEval	GeneticSearch
ConsistencySubsetEval	GreedyStepwise
FilteredAttributeEval	LinearForwardSelection
FilteredSubsetEval	RandomSearch
GainRatioAttributeEval	RankSearch
InfoGainAttributeEval	Ranker
LatentSemanticAnalysis	ScatterSearchV1
OneRAttributeEval	SubsetSizeForwardSelection
PrincipalComponents	
ReliefFAttributeEval	
SVMAttributeEval	
WrapperSubsetEval	

Cuadro 16: Lista de métodos evaluadores y de búsqueda presentes en Weka.

3.4.1.3 *La validación cruzada*

El método utilizado en esta tesis para comparar algoritmos de clasificación es la validación cruzada o *cross-validation*. La validación cruzada divide los datos etiquetados en subconjuntos de entrenamiento y de prueba (test). Los modelos se aprenden sobre los datos de entrenamiento (*training*) y se aplican sobre los datos de prueba (*test*). Los errores de predicción se calculan y promedian para todos los subconjuntos. Así, mediante el método de la validación cruzada, un conjunto D de datos se divide en k partes disjuntas $\{D_1, ..., D_k\}$ procurando que tengan tamaños parecidos y que no se solapen. En cada iteración i (que varía de 1 a k), el algoritmo se entrena con el conjunto $D \setminus D_i$ (conjunto diferencia) y se evalúa en D_i. En la mayoría de los

experimentos de esta tesis he usado *10-fold cross-validation* para estimar la precisión de los algoritmos.

3.4.2 Semantic Orientation CALculator System

Semantic Orientation CALculator System[20] (SO-CAL) [Taboada, Brooke, Tofiloski y col., 2011] es un software destinado a clasificar textos según su polaridad que, al aplicar una aproximación basada en el uso de diccionarios, se utiliza en esta tesis para efectuar la clasificación *no supervisada* de los comentarios sobre productos.

SO-CAL utiliza diccionarios de palabras (adjetivos, adverbios, nombres y verbos) anotadas con su orientación semántica (*semantic orientation*, SO), es decir, polaridad («*good*», «*bad*») e intensidad («*very good*», «*partly good*»). Además, el software incorpora varias estrategias para el tratamiento de los modificadores de la polaridad (*valence shifters*), tanto de intensificadores como de la negación. Como cualquier otra aproximación basada en el léxico, SO-CAL fue diseñado a partir de dos supuestos básicos: 1. las palabras aisladas expresan un orientación semántica de base que es independiente del contexto (*prior polarity*), y 2. la orientación semántica puede representarse de forma cuantitativa.

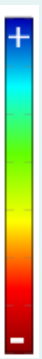
Para la creación del diccionario de adjetivos los autores recuperaron los adjetivos presentes en un corpus de 400 comentarios (279.761 palabras) sobre diferentes tipos de productos obtenidos de la web de Epinions[21]. De manera específica, se recopilaron 25 comentarios positivos y 25 negativos por cada uno de los ocho dominios seleccionados: libros, coches, ordenadores, utensilios de cocina, hoteles, películas, música y auriculares. Cada adjetivo se etiquetó manualmente con su polaridad e intensidad utilizando una escala que va del -5 (adjetivos muy negativos) al +5 (adjetivos muy positivos). Los adjetivos con una polaridad neutra, fueron excluidos del diccionario. El número de entradas que contiene este diccionario es de 2.252 palabras.

SO-CAL también incluye diccionarios independientes para los nombres, los verbos y los adverbios. Los dos primeros diccionarios fueron creados siguiendo las mismas pautas y criterios que el de los adjetivos. El diccionario de nombres contiene 1.142 entradas y el de verbos 903. Por su parte, el diccionario de adverbios, con 745 entradas, se creó de forma automática a partir del diccionario de adjetivos, agregando el sufijo «-ly» al adjetivo correspondiente. Cuando SO-CAL encuentra un adverbio que no aparece en el diccionario, extrae su raíz e intenta identificar un adjetivo que contenga dicha raíz en el diccionario de adjetivos.

La palabras que contienen los diccionarios de nombres, verbos y adverbios se obtuvieron del mismo corpus de opiniones que el de los adjetivos y de dos recursos adicionales: 2.000 comentarios sobre películas del corpus de Pang y L. Lee [Pang y L. Lee, 2004; Pang, L. Lee y Vaithyanathan, 2002] y del diccionario *General Inquirer*[22] [Stone y col., 1966]. Los dos corpus de comentarios representan el registro coloquial, lo que permite incluir palabras como «*ass-kiching*» («zurrar») o «*nifty*» («chulo») en los diccionarios. Al estar elaborado a partir de textos literarios, *General Inquirer* recoge palabras características del registro formal como, por ejemplo, «*adroit*» («diestro») o «*jubilant*» («exultante»). Todos los diccionarios fueron revisados por un comi-

20 El software se ha de solicitar a la Dra. Maite Taboada de la Universidad Simon Fraser.
21 http://www.epinions.com/
22 http://www.wjh.harvard.edu/~inquirer/

té de tres expertos para garantizar su calidad. Un fragmento de cada uno de los diccionarios se puede consultar en el Cuadro 17.



adjetivos	nombres	verbos	adverbios
mega-success 5	masterpiece 5	culminate 4	wondrously 5
priceless 5	perfection 5	exalt 4	delightfully 5
valiant 4	classic 4	exult 4	superbly 4
compelling 4	mastery 4	awe 4	lovably 4
inspirational 3	honour 3	gush 3	extensively 3
capacious 2	quality 2	liberate 2	neatly 2
seductive 1	spirituality 1	attain 1	intimately 1
expensive -1	contradiction -1	argue -1	vastly -1
not_enough -2	hype -2	dump -2	poorly -2
arrogant -3	aggression -3	worsen -3	blindly -3
revolting -4	loathing -4	execrate -4	exploitatively -4
pedantic -4	no_way -4	disgrace -4	idiotically -4
abusive -5	abomination -5	abhor -5	horrendously -5
apocalyptic -5	catastrophe -5	appall -5	agonizingly -5

Cuadro 17: Fragmentos de los cuatro diccionarios que utiliza SO-CAL con la orientación semántica de las palabras.

Aunque la mayoría de las entradas de los diccionarios son palabras simples o aisladas, en SO-CAL también hay entradas compuestas de varias palabras (*multi-words*) que han sido generadas, muchas de ellas, a partir de expresiones regulares: (#run#)_out, (limit)_#NP?#_to, could_[not]?_(care)_less, etc. En general, las expresiones multi-palabra tienen precedencia sobre las palabras aisladas, por lo que si en un texto aparece la expresión «*act funny*» con polaridad -1, ésta polaridad prevalece sobre la del adjetivo aislado «*funny*» que es de +3.

SO-CAL también tiene en cuenta la intensificación y la negación. La intensificación es representada usando dos clases de modificadores de la polaridad: los reductores (*downtoners*), que limitan la fuerza de la orientación semántica (ej. «*slightly*») y los amplificadores, que la incrementan (ej. «*very*»). Tanto la lista de amplificadores como reductores llevan asociado un porcentaje de intensificación (ver Cuadro 18), de manera que la orientación semántica de cada expresión se calcula considerando la orientación semántica de la palabra aislada en relación con su intensificador.

	intensificador	modificación (%)
amplificadores	really	0.15
	very	0.25
	truly	0.25
reductores	somewhat	-0.3
	slightly	-0.5
	less	-1.5
	the_least	-3

Cuadro 18: Porcentaje de intensificación asociado a algunos de los modificadores de la polaridad según SO-CAL.

En cuanto al tratamiento de la negación, SO-CAL asume que es un fenómeno complejo que no puede ser cubierto mediante la simple **inversión** de la polaridad de base de una palabra por efecto de la negación (*switch negation* o *polarity flip*). Por ejemplo, la negación del adjetivo «*excellent*», con una

orientación semántica de base igual a +5, no equivale a -5 («*not excellent*»). De ser así la expresión «*not excellent*» sería equivalente, desde el punto de vista de su orientación semántica, a la de otro adjetivo como «*atrocious*». En todo caso, la polaridad de «*not excellent*» tendría que estar a la par de un adjetivo como «*good*» = +3. Por lo tanto, para tratar la negación, Taboada, Brooke, Tofiloski y col. [2011] utilizan un criterio más analítico basado en el **desplazamiento** de la polaridad de base y que denominan *shift negation*. Este criterio consiste en cambiar la polaridad de base de una palabra a partir de un valor fijo que, en esta versión del SO-CAL, equivale a 4. Así, la negación de dos adjetivos como «*excellent*» y «*atrocious*», se establece como sigue:

«*not excellent*»: $+5 - 4 = \mathbf{+1}$
«*not atrocious*»: $-5 + 4 = \mathbf{-1}$

Por último, SO-CAL omite ciertos fragmentos del texto que contienen palabras que indican que la polaridad de la oración en la cual aparecen no son fiables debido a que el evento descrito en la oración no ha ocurrido (*irrealis blocking*). Por ejemplo, en la oración «*You'd expect such a basic concept to be implemented correctly*», la palabra «*expect*» indica que el evento «*implemented*» no ha tenido lugar. Algunas de las marcas que se utilizan para detectar estos fragmentos son: modales, condicionales, elementos de polaridad negativa (*negative polarity items*, NPIs) como «*any*» o «*anything*», ciertos verbos («*expect*», «*doubt*»), preguntas, etc.

3.5 CONCLUSIONES

En este capítulo he presentado las herramientas y los recursos que aplico para el tratamiento computacional de la polaridad de los comentarios mediante el análisis de su estructura discursiva. Las principales características de estas herramientas y recursos son:

- La fácil obtención: la mayor parte del software utilizado es de libre distribución;
- La fácil integración: los corpus, léxicos y ontologías usan XML (y tecnologías asociadas) como lenguaje de etiquetado;
- La simplicidad: son herramientas que no exigen un entrenamiento previo, que no consumen muchos recursos del sistema y que corren bajo diferentes sistemas operativos;
- La calidad: son recursos robustos y eficientes, desarrollados por centros consolidados de investigación.

4 CARACTERIZACIÓN DEL GÉNERO DISCURSIVO DE LOS COMENTARIOS

Resumen

En este capítulo analizo la estructura discursiva y el registro lingüístico de los comentarios. El objetivo de ambos análisis es determinar si los comentarios sobre productos constituyen un género discursivo propio como paso previo al análisis de la polaridad. En primer lugar, el análisis de la estructura discursiva busca identificar regularidades en el tipo y la distribución de los diferentes segmentos que componen los comentarios. Para ello, efectúo una propuesta de segmentación que cotejo con un corpus real de comentarios sobre hoteles. En segundo lugar, el análisis del registro lingüístico busca identificar regularidades léxicas y morfosintácticas entre estos textos de opinión. Este análisis lo llevo a cabo mediante dos tipos de experimentos: un primer grupo de experimentos están orientados a contrastar el registro lingüístico de los comentarios entre sí y un segundo grupo, entre los comentarios y textos periodísticos. Los resultados obtenidos en esta parte de la investigación aparecen publicados en:

- John Roberto, Maria Salamó y M. Antònia Martí [2015a], «Genre-Based Stages Classification for Polarity Analysis», en *The 28th Florida Artificial Intelligence Society Conference (FLAIRS), USA*, vol. 1, pág. 1-6
- John Roberto, Maria Salamó y M. Antònia Martí [2013], «Clasificación automática del registro lingüístico en textos del español: un análisis contrastivo», *Linguamática*, 5, 1, pág. 59-67
- John Roberto, Maria Salamó y M. Antònia Martí [2012], «Análisis de la riqueza léxica en el contexto de la clasificación de atributos demográficos latentes», *Procesamiento de Lenguaje Natural*, 1, 48, pág. 97-104
- John Roberto, M. Antònia Martí y Paolo Rosso [2011], «Sistemas de Recomendación basados en Lenguaje Natural: Opiniones vs. Valoraciones», *Actas IV Jornadas Tratamiento de la Información Multilingüe y Multimodal (TIMM)*, pág. 45-48

4.1 INTRODUCCIÓN

Un comentario es un evento comunicativo reconocido social y culturalmente que consiste en el acto de valorar positiva o negativamente un producto, bien o servicio. Desde las diferentes disciplinas que tienen por objeto el análisis de los medios de comunicación social para el consumo (*Consumer Generated Media*), se asume que los comentarios son un tipo específico de texto asociado a la expresión de opiniones en la Web. Los usuarios de estos medios se valen de métodos de trabajo grupal (conducta colaborativa) para mejorar la toma de decisiones en el ámbito comercial. Dentro de los recursos o servicios que se utilizan para generar este tipo de contenidos tenemos los foros, blogs o microblogs como Twitter[1] y sitios web especiali-

1 https://twitter.com

zados (*consumer review sites*) como Epinions[2], Yelp[3], Amazon[4], RateItAll[5] o Tripadvisor[6].

El objetivo de este capítulo es demostrar que los comentarios sobre productos constituyen una clase específica de texto que podemos vincular a un género discursivo concreto. Siguiendo los postulados de la Lingüística Sistémico-Funcional, en esta tesis asumo que los comentarios constituyen un género discursivo propio o autónomo **si** poseen una estructura discursiva estable y **si** comparten un mismo registro lingüístico. La Lingüística Sistémico-Funcional [Martin, 2001] define los géneros discursivos como clases estables de textos que se caracterizan por *(i)* presentar unas propiedades estructurales recurrentes y *(ii)* compartir un mismo registro lingüístico. En primer lugar, el análisis de la estructura permite conocer si los comentarios comparten unos determinados tipos de segmentos asociados a un **propósito comunicativo** concreto. En segundo lugar, el análisis del registro permite saber si los comentarios presentan similitudes léxicas y morfosintácticas significativas determinadas por el hecho de configurar una misma **situación comunicativa**. La Figura 3, basada en el trabajo de Martin [2001], señala que para caracterizar los comentarios como un género discursivo es indispensable identificar su propósito comunicativo (estructura) y su situación comunicativa (registro).

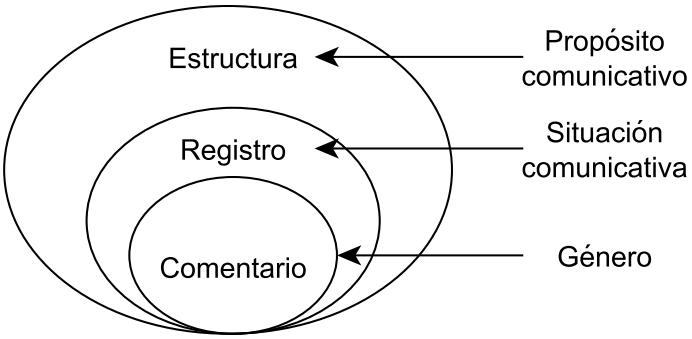


Figura 3: Relación entre estructura y registro aplicada a la caracterización de los comentarios como género discursivo (basado en Martin [2001]).

En la Sección 4.2 analizo la estructura discursiva de los comentarios sobre productos. En primer lugar, defino los principales tipos de segmentos que conforman estos textos de opinión basándome en la descripción de su propósito comunicativo. En segundo lugar, analizo la distribución de estos segmentos en un corpus de 150 comentarios sobre hoteles. En la Sección 4.3 analizo el registro lingüístico de los comentarios sobre productos. Para ello, presento varios experimentos orientados a determinar si los comentarios comparten el mismo registro lingüístico. Un primer grupo de experimentos contrastan el registro lingüístico de los comentarios entre sí (análisis intra-textual) y un segundo grupo de experimentos contrastan el registro lingüístico de los comentarios con el de otra clase de textos (análisis inter-textual). Por último, en la Sección 4.4 presento las conclusiones generales del capítulo.

2 http://www.epinions.es
3 http://www.yelp.es
4 http://www.amazon.es
5 http://www.rateitall.com
6 http://www.tripadvisor.es/

4.2 ANÁLISIS DE LA ESTRUCTURA DISCURSIVA

En esta sección presento el análisis de la estructura discursiva de los comentarios sobre productos. En la primera parte del análisis expongo la propuesta para la segmentar los comentarios. Dicha propuesta se basa en las nociones de género discursivo y tipología textual. En la segunda parte del análisis proporciono una definición de cada uno de los tipos de segmentos. En la tercera parte del análisis presento un estudio que detalla la forma en que estos tipos de segmentos se distribuyen en un corpus real de comentarios sobre hoteles.

4.2.1 Propuesta de segmentación

En esta tesis considero que los comentarios sobre productos se componen de una serie de segmentos que contribuyen de manera diferente a la asignación de la polaridad del texto. La selección de estos segmentos no es aleatoria sino que está motivada por los géneros discursivos y las tipologías textuales, dos ámbitos que determinan las regularidades estructurales en el uso del lenguaje.

Los géneros discursivos

Autores como Eggins [2004], Stenstrom [1994], Ricci y Wietsma [2006], Bieler y col. [2007], Taboada, Brooke y Stede [2009] y Taboada [2011], han desarrollado propuestas para segmentar diferentes clases de textos de opinión. Así, Eggins [2004] y Stenstrom [1994] definen la estructura discursiva de un texto como una totalidad en la que las partes se relacionan unas con otras mediante criterios formales y funcionales. Para los autores, cualquier género discursivo ha de poder dividirse, como mínimo, en tres segmentos básicos que presentan, desarrollan y concluyan la exposición de un tema. En esta misma línea se expresan Ricci y Wietsma [2006] quienes abogan por una segmentación tripartita de los comentarios sobre productos. Para estos autores, el comentario es una pieza de texto subjetiva que describe las **experiencias**, el **conocimiento** y las **opiniones** que un usuario tiene de un producto[7] [Ricci y Wietsma, 2006, p. 297].

Bieler y col. [2007] y Taboada, Brooke y Stede [2009] proponen dividir los comentarios en «zonas» formales y funcionales. Según Bieler y col., las zonas formales son partes del comentario que aportan datos o información factual como por ejemplo, en el caso de los comentarios sobre películas, el título o el nombre del director. Las zonas funcionales son partes del comentario que presentan el contenido desde el punto de vista subjetivo (*comment*) u objetivo (*descriptive*). En un trabajo posterior, Taboada [2011] propone clasificar los diferentes segmentos que componen los comentarios sobre películas según la «fórmula» presentada en (1). La característica principal de esta descripción del género es que, siguiendo las ideas de Hasan [1996] respecto al «potencial de estructura de género», Taboada considera los segmentos como «potenciales» (opcionales) o «realizados» (necesarios). Adicionalmente, la fórmula (1) tiene en cuenta el orden de aparición de los segmentos, una propiedad fundamental para definir las formas prototípicas del género.

(1) (Subject matter) $\wedge$ (Background)/Evaluation $\wedge$ (Subject matter)/(Plot)/(Background) $\wedge$ Evaluation $\wedge$ (Characters) $\wedge$ Evaluation

7 Además de una valoración final expresada mediante algún valor numérico.

(Asunto) ∧ (Antecedentes)/Evaluación ∧ (Asunto)/(Trama)/(Antecedentes) ∧ Evaluación ∧ (Personajes) ∧ Evaluación

Ganu y col. [2010] han analizado la estructura de los comentarios sobre restaurantes como parte del proyecto URSA[8] (*User Review Structure Analysis*). Su propuesta para la segmentación de comentarios presenta un carácter marcadamente funcional ya que efectúa la segmentación a partir de la identificación de los diferentes temas que se tratan en el comentario (ej. tipo de comida o calidad del servicio). No obstante, también han considerado algunos segmentos de carácter formal (ej. título o puntuación).

En términos generales, la clasificación de los segmentos que componen los comentarios sobre productos se ha realizado teniendo en cuenta el contenido más que la función comunicativa de los segmentos. Esto lleva a una dependencia importante del dominio. Por ejemplo, los segmentos «trama» o «personajes» solo son útiles para describir los comentarios sobre películas. Considero, por tanto, que para efectuar una descripción extendida de los tipos de segmentos se ha de atender también a criterios generalistas para la clasificación de textos como el de las tipologías textuales.

Las tipologías textuales

Los géneros discursivos se construyen a partir de diferentes tipos textuales. La tipología textual está relacionada con la selección de criterios para la clasificación de textos. A diferencia de los géneros discursivos que son definidos en base a criterios no lingüísticos, los diferentes tipos textuales han sido definidos siguiendo determinados categorías cognitivas o criterios lingüísticos[9]. El resultado es un repertorio cerrado de formas entre las que figuran la narración, la descripción y la argumentación como tipos textuales básicos:

NARRACIÓN: Es un tipo textual relacionado con la contextualización de los eventos en el tiempo. Su función consiste en informar sobre acciones y hechos. Ochs [1997] afirma que la narración tiene una estructura determinada que la distingue de los otros tipos discursivos y es un medio para que los miembros de una comunidad representen los eventos, sus pensamientos, sus sentimientos y sus creencias. La esencia de la narración es el cambio temporal o cronológico que tiene como elemento discursivo básico, el «**evento**». Por lo tanto, es lógico definir la narración como una representación secuencial de eventos, reales o ficticios que componen un texto. En este sentido, según Labov [1972], todos los textos narrativos tienen en común la descripción de una transición temporal de un estado de cosas a otro. Para éste y otros investigadores como Reinhart [1984], desde el punto de vista lingüístico, las narraciones deben contener al menos dos cláusulas narrativas ordenadas temporalmente. El cambio de orden de estas cláusulas puede modificar la interpretación semántica de la secuencia temporal original.

DESCRIPCIÓN: Es un tipo textual relacionado con la percepción de los objetos y sus propiedades [Loureda Lamas, 2009]. Su función consiste en informar sobre el estado de las cosas, fenómenos y situaciones. En esencia, se trata de un tipo de texto con una función claramente representativa en el que prevalece la referencia al objeto. En este sentido, el

8 `http://spidr-ursa.rutgers.edu/`

9 Mientras que los géneros discursivos son **productos socioculturales**, es decir, formas discursivas convencionales con un valor funcional, los tipos textuales son **realidades abstractas** con valor nominal que forman parte de un modelo conceptual específico: modelo de Grosse [Grosse, 1976], modelo de Werlich [Werlich, 1976], modelo secuencial de Adam [Adam, 2011], etc.

tema es el elemento predominante ya que supone el punto en el que se articula toda la exposición. En cuanto a su estructura, es de tipo piramidal, es decir, que permite progresar a partir de la tematización[10] y el énfasis. La «**aspectualización**», entendida como la enumeración de partes o propiedades de las que consta el tema y su puesta en relación (ej. comparación), caracteriza este tipo de texto. Si se tiene en cuenta la actitud del emisor, la descripción puede presentar dos variantes: descripción objetiva y descripción subjetiva[11].

ARGUMENTACIÓN / VALORACIÓN: La argumentación es un tipo textual relacionado con la expresión de juicios de valor. Su función consiste en expresar opiniones, defender o rechazar una idea y, en general, persuadir. La argumentación se basa en la evaluación de las relaciones de contraste entre conceptos. Dentro de este tipo textual se suele incorporar la valoración como un tipo especial de argumentación centrada en la función de «**automanifestación**» o *self-expression* [Grosse, 1976]. El tipo textual valorativo, además de argumentar, tiene una fuerte inclinación a la persuasión, es decir, a regular el comportamiento del destinatario, de ahí que aparezca en los estudios discursivos relacionados con la expresión pública de opiniones (ej. artículo periodístico).

La principal propiedad de estos tipos textuales es que se organizan de manera secuencial y jerarquizada en el interior de un texto. Lo normal es que no existan textos «puros» –puramente narrativos, puramente descriptivos, etc.– sino que predomine un tipo sobre los demás, que es siempre reconocible. Hatim e I. Mason [1990] denominan esta propiedad como «hibridación». Algunos autores como Petitjean [1987, p. 81] y Roulet [1991, p. 126] asignan a las secuencias descriptivas un papel secundario en el sentido de que se suelen emplear como elemento de apoyo en textos predominantemente narrativos o argumentativos. Por el contrario, de la narración se ha llegado a afirmar que es el tipo cultural y universal más común. En este sentido se manifiesta Barrera para quien «por encima de otros órdenes tales como la descripción, la exposición, la argumentación y la instrucción, estaría el texto narrativo como la forma expresiva más relevante de la especie» [Barrera, 1995, p. 11]. De la argumentación, Verschueren [1999, p. 46] ha afirmado que es «la fuerza organizativa básica que subyace a toda la comunicación lingüística».

La conclusión general que se puede extraer de los estudios sobre el género discursivo y las tipologías textuales es que el género de los comentarios se puede describir en términos de un número reducido de segmentos. Por lo tanto, propongo caracterizar los comentarios sobre productos a partir de tres tipos de segmentos: NARRATIVO, DESCRIPTIVO y VALORATIVO. Esta estructura tripartita está motivada por tres de los tipos textuales básicos descritos anteriormente: narración, descripción y argumentación. Estos segmentos están orientados a expresar las experiencias, el conocimiento y la opinión que un usuario tiene de un producto. En el siguiente apartado defino cada uno de estos tipos de segmentos.

10 La tematización es un mecanismo sintáctico en virtud del cual el tema (normalmente la unidad que desempeña la función de sujeto) se ubica en un lugar periférico de la oración, que suele coincidir con la posición inicial.

11 La descripción objetiva se ajusta al criterio general y a la realidad (ej. «el hotel está bien localizado») mientras que la descripción subjetiva se ajusta más al criterio individual («la habitaciones están bien decoradas»).

4.2.2 Definición de los tipos de segmentos

Este apartado tiene por objeto desarrollar la propuesta de segmentación de los comentarios sobre productos presentada en el apartado anterior. Para ello, proporciono una definición de los tres tipos básicos de segmentos que conformarían un comentario (narración, descripción y valoración) y que aparecen identificados en el texto del Cuadro 19. La notación usada para identificar los tres tipos de segmentos se mantendrá a lo largo de toda la tesis.

MI COCHE ★ ★ ★ ★ ★

Aún me acuerdo, cuando salió el primer anuncio, basado en la película Abierto hasta el amanecer. Quedé prendado de ese coche. Yo era estudiante, como lo iba comprar. Nunca lo imaginé, hasta que acabé de estudiar y pude comprarlo. Me enamoró, bonito, potente, estaba bastante bien equipado, y hasta el día de hoy no tengo ninguna queja. Sí, es un coche pequeño, pero yo buscaba eso, que tuviera algo de potencia y de eso tiene bastante. El consumo no es alto, depende de lo que le pises, pero si vas a una marcha normal no hay ningún problema. Es un coche hecho para gente joven, para una pareja sin hijos, es cómodo para largos viajes. Para mí no tiene ningún defecto. Se lo recomiendo.

segmento narrativo segmento descriptivo segmento valorativo

Cuadro 19: Propuesta de segmentación de los comentarios.

4.2.2.1 *El segmento narrativo*

Parte de la información que compone un comentario hace referencia a un conjunto de EVENTOS que suelen ir asociados a la valoración del producto. Por ejemplo, los fragmentos (2a) al (2c) describen una serie de situaciones (narraciones) relacionadas con la valoración de un coche (2a), un hotel (2b) y unos auriculares (2c). La principal función comunicativa del segmento narrativo es **contextualizar el comentario**, es decir, aportar información que facilite su comprensión. Por lo tanto, desde el punto de vista formal, los segmentos narrativos se caracterizan por contener expresiones temporales (ej. «*in 1992*», «*before*», «un día») y, al narrar vivencias personales, suelen ir en primera persona (ej. «*My dad*», «Mi esposa», «Yo tenía»). Adicionalmente, como se aprecia en los ejemplos, el segmento narrativo aporta información personal sobre el usuario, especialmente información demográfica (edad, sexo, estado civil, etc.): «*my graduation in 1992*» (2a), «Mi esposa ... amigos en el Reino Unido» (2b).

(2) a. *My dad bought me a Saturn for my graduation in 1992 before all the marketing hype.*

b. Mi esposa y yo tuvimos ocasión de pasar dos noches en Londres en un viaje para visitar a amigos en el Reino Unido.

c. Yo tenía unos Panasonic Old Style rojos monísimos. Me habían costado 50 euros, se oían de maravilla y me encantaban, pero se rompieron, y tuve que buscar unos sustitutos de emergencia porque la verdad es que los usaba mucho y no tenía otros para sustituirlo. Y un día, en el Carrefour, vi los Philips SHP1900...

4.2.2.2 *El segmento descriptivo*

Otro tipo de información que forma parte de un comentario tiene que ver con las diferentes propiedades, componentes, aspectos o CARACTERÍSTICAS que integran el producto que es objeto de valoración (conocimiento del producto). Por ejemplo, el fragmento (3a) describe y valora la ubicación, la habitación, el baño, el desayuno, la atención y las áreas comunes de un hotel. De la misma forma, en el ejemplo (3d) se describen, entre otras características, la suspensión (*suspension*), el diseño (*design*) y el limpiaparabrisas (*windscreen washers*) del Volkswagen Beetle 1300. La principal función comunicativa del segmento descriptivo es **presentar las características del producto**, tanto desde un punto de vista objetivo (ejemplo 3b) como subjetivo (ejemplo 3c). Desde el punto de vista formal, los segmentos descriptivos se caracterizan por contener una alta presencia de sustantivos (ej. «habitación», «ducha», «*suspension*», «*models*»), así como de adjetivos y adverbios (ej. «pequeña», «abundante», «amplio», «mejora», «*heavy*», «*stupid*»). Como se puede observar en los ejemplos, este tipo de segmento aporta información detallada sobre el producto: «desayuno completo y abundante» (3a), «son habitaciones de unos 30 metros cuadrados» (3b), «su imagen exterior es buena» (3c), «*the windscreen washers used a pressurised bottle*» (3d). Puesto que esta clase de información se centra en el producto antes que en el usuario, el contenido de los segmentos descriptivos depende del dominio y, más exactamente, del conocimiento que el usuario tiene del dominio. La información que aporta el segmento descriptivo varía en función del conocimiento y la experiencia que tiene el usuario con determinado producto. Según Alba y Hutchinson [1987], esta información «se basa en el aprendizaje y la experiencia que ha tenido el usuario con otras marcas y productos similares», de ahí que la comparación y el contraste sean elementos característicos de este tipo de segmentos: «sin ser tan amplio como una berlina» (3c), «*the design was based on a tn tray with a roof*» (3d).

(3) a. Hotel excelentemente ubicado frente al metro y a pocas cuadras de la costa y rambla. Habitación moderna algo pequeña con buen baño. Desayuno completo y abundante. Cálida atención. Áreas comunes algo anticuadas.

b. Son habitaciones de unos 30 metros cuadrados, sin aire acondicionado ni calefacción. No hay WC ni ducha, solamente una pica con un grifo.

c. Su imagen exterior es buena sin arriesgar en exceso. El espacio interior bueno, sin ser tan amplio como una berlina pero mejora ante su hermano el 3p/5p ...

d. *Suspension was provided by 'torsion bars', rather than coil springs (although later models featured a now traditional 'McPherson Strut'. Rather than being built around a monocoque 'steel box' the design was based on a 'tn tray with a roof' – heavy, but solid, and very forgiving as long as you didn't let the sills rust. There were several ingenious touches which were both brilliant, and stupid at the same time. The windscreen washers used a pressurised bottle, earlier models charged this from the spare tyre, later this had it's own valve, which needed to be charged with a footpump or on a garage forecourt.*

4.2.2.3 *El segmento valorativo*

Otra parte de la información que conforma un comentario, la más característica, expresa la actitud personal del usuario respecto del producto sobre el cual esta opinando (el producto en su totalidad). Por ejemplo, los fragmentos (4a) al (4g) especifican cuál en la postura general de los usuarios en relación con el hotel Patio Brancion (4a, 4b y 4c), los auriculares MonsterCable Beats

(4d, 4e y 4f) y el Volkswagen Golf 1.4 FSI (4g). La principal función comunicativa del segmento valorativo es **persuadir al lector para que adquiera o rechace un producto**, por lo que este tipo de segmento tiene un claro propósito persuasivo. La valoración suele presentarse como una reflexión personal del usuario, quien se muestra en favor o en contra del producto (4d).

Adicionalmente, el segmento valorativo sirve para sintetizar la opinión expresada en el comentario. Por lo tanto, desde el punto de vista formal, los segmentos valorativos se caracterizan por contener expresiones de recomendación (ej. «no me atrevo a recomendarlo», «*beware before you purchase any new or used golf*») y/o conclusión (ej. «en fin», «en conclusión», «en general», «*so*»). Como se aprecia en los ejemplos, este tipo de segmento enlaza los dos anteriores (caracterización de usuario y caracterización del producto) poniendo en relación directa el usuario y el producto. Esta información es útil ya que no siempre la descripción de las características del producto refleja la opinión del usuario respecto de aquel. Por ejemplo, en (5) tenemos un comentario sobre La Taberna Española en el que se describe la comida, el personal, la localización y el precio de forma positiva pero, contrariamente, la valoración final del usuario es negativa.

(4) a. En fin, para dormir y desayunar exclusivamente y no es que esté mal, pero no me atrevo a recomendarlo y, por supuesto, no creo que vuelva.
b. En general, estuvimos muy contentos con este hotel.
c. *So I would recommend this hotel to everyone who comes to Paris ... really worth that money.*
d. Desde mi punto de vista son muy buenos yo los tengo hace como dos semanas y no paro de escucharlos.
e. En conclusión, los recomiendo para escuchar música o ver películas en el iPhone/iPod durante tus desplazamientos y poder responder a una llamada entrante con solo tocar un botón.
f. Para todo aquel que no los conozca o esté pensando en comprar unos auriculares y no se decida por una marca u otra, yo los recomiendo totalmente.
g. *So beware before you purchase any new or used golf in the future it won't affect me because I won't buy another.*

(5) Estaba dando una vuelta, buscando una terraza donde sentarme con mi hermana a comer algo, y me apetecía probar un sitio nuevo, empezamos con 3 entrantes de los cuales las bravas se llevan la guinda, muy ricas!! De Segundo, la paella no estaba muy allá la verdad, pero el cordero estaba bien, y el postre no estaba mal. Trato del personal excelente, nos atendió un camarero muy amable, sitio un poco turístico, pero para serlo no está mal de precio. Sin embargo yo no volvería.

En la siguiente sección, analizo la distribución de estos tres tipos de segmentos a partir de un corpus real. Mi objetivo es verificar que los comentarios presentan una distribución estable y uniforme en cuanto al uso de los segmentos descritos anteriormente. Esta es una condición indispensable para clasificar estos textos de opinión en un mismo género discursivo.

4.2.3 Distribución de los tipos de segmentos

En esta sección presento los resultados obtenidos al analizar la distribución de los segmentos narrativo (`N`), descriptivo (`D`) y valorativo (`V`) en un conjunto de 150 comentarios extraídos del corpus HOpinion. La anotación de los 150 comentarios del corpus HOpinion se realizó de forma manual[12]

12 La anotación automática se presenta en la Sección 5.2.

siguiendo los criterios funcionales descritos en el apartado anterior. Los resultados de la anotación se pueden ver en las Figuras 4, 5 y 6.

En la Figura 4 he agrupado los comentarios que, tras la anotación manual de cada uno de los segmentos, comparten idéntico patrón estructural. En total he identificado 27 formas diferentes de organizar los segmentos. La mayoría de los comentarios se agrupan en torno al patrón: *narración – descripción – valoración*, que puede ser caracterizado como una secuencia de segmentos (o «movimientos», en palabras de Swales [1990]) del tipo `N – D – V`. Este patrón representa la manera «prototípica» de estructurar los comentarios sobre productos por dos razones básicas. En primer lugar, está su elevada frecuencia de aparición: prácticamente el 30 % de los comentarios (44 textos) se construyen según esta secuencia. En segundo lugar, es el patrón más coherente de las 27 representaciones identificadas puesto que presenta la información de forma incremental: presentación y contextualización del comentario, descripción de las características del producto y, por último, valoración general del producto. Si eliminamos el segmento narrativo de este patrón estructural obtenemos el segundo patrón más recurrente: `D – V` (17 %). Los patrones menos frecuentes están formados por secuencias complejas, por ejemplo, `N – D – V – N – D – N` (0.6 %). La baja frecuencia de estas secuencias se explica porque la complejidad es una propiedad incompatible con el principio de relevancia comunicativa que caracteriza cualquier género discursivo[13].

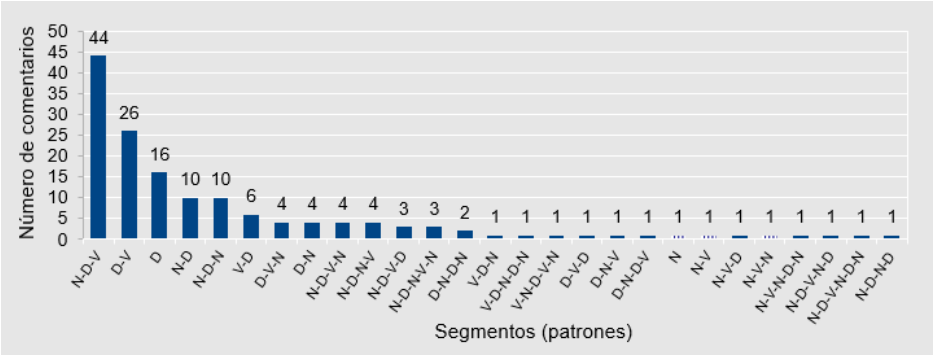


Figura 4: Comentarios agrupados según su patrón estructural.

El segmento descriptivo es el más frecuente y el más irregular de los tres. Es el más frecuente puesto que aparece en prácticamente todas las configuraciones. En total, 147 de los 150 comentarios (98 %) poseen dicho segmento. Esto se debe a que los usuarios normalmente describen los *pros* y los *contras* para justificar sus valoraciones. Las descripciones son un elemento que diferencia los comentarios de otros géneros de opinión aún más breves como, por ejemplo, los *tweets*. El segmento descriptivo es el que presenta una distribución menos regular puesto que aparece en cualquier posición dentro del comentario. De los 147 comentarios con segmento descriptivo, 78 veces aparece en una posición intermedia (53 %)[14], en 39 al principio (26 %) y 23

13 De acuerdo con Wilson y Sperber [2004] y Sperber y Wilson [1986], un enunciado es relevante cuando produce el máximo efecto en el interlocutor con el menor esfuerzo cognitivo.

14 La suma de estos porcentajes no equivale al 100 % puesto que para calcularlos se han tenido en consideración los segmentos que aparecen en varias posiciones. Por ejemplo, en la secuencia `D–V–D`, se contabilizan dos apariciones del segmento descriptivo, una al principio y otra al final del comentario.

veces al final (15 %). En 16 casos (10 %) la descripción aparece como único segmento (ver ejemplo 6).

(6)  Hotel de 3 estrellas decente, pero al llegar nos dieron la habitacion de minusvalidos, con un baño sin cortinas, y al pedir que la cambiaran, nos dieron otra en fumadores que la verdad olia bastante a tabaco. En general la habitación esta bien (baño normalisimo) y el desayuno completo. Ojo, la calefacción y el aire acondicionado te la dan desde recepción, por lo que no tienes opción de escoger temperatura e intensidad.

El segmento valorativo es el segundo en frecuencia. En total, 106 de los 150 comentarios poseen por lo menos un segmento de tipo valorativo (70 %). Este segmento es altamente regular. De los 107 comentarios con segmento valorativo, 77 veces se ubica al final del comentario (72 %)[15], 21 veces en medio (20 %) y sólo 9 veces aparece al principio (8 %). En el ejemplo (7) presento un caso de esta última e inusual situación.

(7)  Aparecerá un 1 en nuestra puntuación, porque es imposible ponerle un 0, que es lo que merece este hotel. Las habitaciones son de lo peor que nunca nos encontramos en una larga lista de viajes, las puertas, ventanas y camas están pidiendo a gritos ser cambiadas, el resto del hotel te invita a pasar el menor tiempo posible allí...

Los segmentos valorativos diferencian los comentarios sobre productos de otros géneros próximos, como sería el caso de las críticas especializadas. Efectivamente, en el ejemplo (8) aparece una crítica de cine extraída de una columna de opinión cinematográfica[16] en la que se puede apreciar la falta de cualquier valoración personal de la película efectuada en primera persona. Esto es porque las críticas, a diferencia de los comentarios, tienden a ser más objetivas y en ellas predominan las descripciones.

(8) «*El chico del millón de dólares*»

Mira aquí un momento?

Escrito por José Arce el 24.10.14 a las 17:39

Nuevo ejemplo de colonialismo cinematográfico que clama las bondades de la tierra de las oportunidades. Aspecto telefílmico, mal planteamiento de la figura central y sus acompañantes, un error global.

El segmento narrativo es el menos frecuente de los tres tipos de segmentos que componen el género de los comentarios. Este segmento está presente en 101 de los 150 comentarios (67 %). Los usuarios pueden omitir este tipo de segmento puesto que las circunstancias que rodean la opinión no son esenciales para cumplir con el objetivo del comentario. Este segmento también es altamente regular. De los 101 comentarios con segmento narrativo, 85 veces se ubica al principio (84 %)[17], 30 veces al final (30 %) y 17 veces en medio (16 %). Solo un comentario se compone exclusivamente de un único segmento narrativo (ver ejemplo 9). Tanto por la forma como por el contenido y la función comunicativa, el ejemplo (9) no constituye un verdadero comentario. Los comentarios sobre productos deben presentar un segmento descriptivo o, en su defecto, un segmento valorativo. Ambos tipos de segmentos se inclu-

15 Ver nota 14 a pie de página.
16 `www.labutaca.net`
17 Ver nota 14 a pie de página.

yen en las propuestas de Bieler y col. y Taboada, Brooke y Stede, reseñadas en el Apartado 4.2.1.

(9)  Yo no he estado alojada, pero no me ha hecho falta. Fuimos a una especie de congreso de bodas que hacían en este hotel. Estando allí vemos que todo el mundo sale a la calle y que una chica que había vestida de novia va a tirar el ramo. Cuando salimos, los organizadores dicen que el que coja el ramo le tocan 2 noches de hotel en la suite presidencial del hotel para el día de la boda [...] luego vienen las desilusiones y eso no es lo que quiero para el día de mi boda...

En la Figura 5 he agrupado los comentarios según el número de segmentos que contienen. Según el análisis, el 85 % de los comentarios (127 textos) se construyen a partir de secuencias simples, es decir, de 1 a 3 segmentos. Más aún, el 42 % de los comentarios (63 textos) se componen de tres segmentos. Sólo el 15 % de los comentarios se construyen a partir de secuencias complejas de cuatro (15 textos), cinco (7 textos) o seis (1 texto) segmentos. Adicionalmente, el porcentaje de repetición de un mismo tipo de segmentos se puede calcular mediante la fórmula $(S_r * 100)/R_s$. En la fórmula, S_r es el número de comentarios en los que el segmento s ocurre más de una vez y R_s es el número total de comentarios que contienen el segmento s. Así, 0.9 % es el porcentaje de repetición del segmento valorativo, 7.4 % el del descriptivo y 29.7 % el del narrativo. El segmento valorativo es, por lo tanto, el que menos se repite en un mismo comentario.

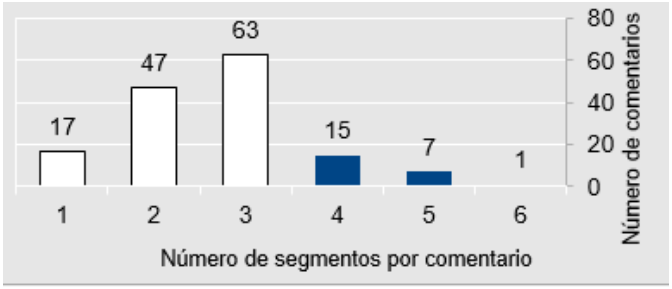


Figura 5: Comentarios agrupados según el número de segmentos que contienen.

Finalmente, en la Figura 6 he agrupado los comentarios según las distintas combinaciones que se pueden establecer a partir de los tres tipos de segmentos. Según los resultados, las combinaciones más usadas son `N – D` y `D – V`. La primera combinación está presente en 90 comentarios y la segunda en 86. Ambas combinaciones se caracterizan por poseer un segmento descriptivo. El segmento narrativo, por el contrario, forma parte de las combinaciones menos frecuentes (excepto, evidentemente, `N – D`). La elevada frecuencia del patrón `N – D` refuerza la idea de que los segmentos narrativos sirven para introducir o contextualizar los segmentos descriptivos, de ahí que aparezcan antes. Por su parte, el patrón `D – V` indica que los segmentos valorativos sintetizan y refuerzan los descriptivos, de ahí que vayan después.

La distribución prototípica de estos segmentos queda reflejada en la fórmula (10) donde «∧» indica secuencialidad, «( )» opcionalidad, «[ ]» obligatoriedad y «*» que el segmento se puede repetir.

(10) (Narración) ∧ [Descripción* ∧ (Valoración)]

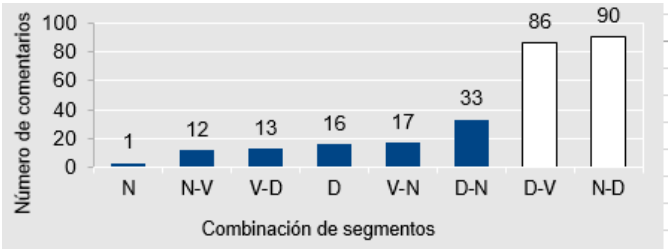


Figura 6: Comentarios agrupados según su proximidad.

4.3 ANÁLISIS DEL REGISTRO LINGÜÍSTICO

El registro lingüístico se refiere al uso de determinadas variedades de la lengua asociadas a situaciones comunicativas concretas. De acuerdo con Biber [1992], la co-ocurrencia de rasgos lingüísticos es fundamental para la caracterización del registro. En esta sección presento dos tipos de análisis orientados a determinar si los comentarios sobre productos comparten el mismo registro lingüístico:

ANÁLISIS INTRA-TEXTUAL: Los comentarios son agrupados según diferentes criterios demográficos (edad, sexo y procedencia de los autores) para determinar si existen diferencias léxicas y/o morfosintácticas significativas entre estos conjuntos de textos.

ANÁLISIS INTER-TEXTUAL: Los comentarios son contrastados con otra clase de textos (artículos periodísticos) para determinar si existen diferencias léxicas y/o morfosintácticas importantes entre ambos conjuntos de textos.

Dado que autores como Sharoff y col. [2010] argumentan que la información léxica es más efectiva para la clasificación de textos, en los análisis intra-textual e inter-textual aplico diversas métricas de la riqueza léxica. Los índices de riqueza léxica son un medio efectivo para el estudio cuantitativo de los textos. Ménard [1983, p. 16] señala que la riqueza léxica ha de entenderse como «un lugar de comparación entre dos o más textos en función del número y del tipo de palabras que figura en cada uno de ellos». En el siguiente apartado defino la métricas de riqueza léxica que aplicaré para determinar el registro lingüístico de los comentarios mediante los análisis intra e inter-textual.

4.3.1 Las métricas de riqueza léxica

La riqueza léxica (`RiqLex`) evalúa los textos en términos de la variedad y la cantidad del vocabulario que contienen de acuerdo con la relación types/tokens por lo que su cálculo tiene un coste computacional muy bajo y, además, es totalmente independiente del dominio.

La `RiqLex` [Read, 2005] se compone de tres dimensiones que son la densidad (`DenLex`), la sofisticación (`SofLex`) y la variación (`VarLex`)[18]. La `DenLex`

18 Adicionalmente, Laufer y Nation [1995] reconocen una dimensión más, la originalidad léxica. No obstante, el índice de originalidad léxica (`OriLex`) es una medida individual: se emplea para comparar la producción escrita de una persona con relación a un grupo. La `OriLex` calcula el porcentaje de palabras que usa la persona y que no son usadas por nadie más dentro de ese mismo grupo. Por este motivo no la empleo en esta investigación.

[Ure, 1971] calcula la proporción entre el número de palabras léxicas y el número total de palabras (léxicas y gramaticales) de un texto. La `SofLex` [Read, 2005], también conocida como singularidad léxica, mide la proporción de palabras «sofisticadas» o «avanzadas» presentes en un texto. Finalmente, la `VarLex` [Granger y Wynne, 2000], denominada diversidad léxica [Malvern y col., 2004] o ámbito léxico [Crystal, 1982], mide el número de *types* o palabras diferentes que hay en un texto.

En el Cuadro 20 presento el conjunto de métricas seleccionadas. En total hay una medida de la `DenLex` (1), cinco de la `SofLex` (2-6) y 22 de la `VarLex` (7-29). El número real de métricas es de 32 pues en las fórmulas 16, 20, 25 y 29 el subíndice $_G$ puede corresponder con cualquiera de las categorías de base léxica: n (nombre), v (verbo), a (adjetivo) o r (adverbio). Adicionalmente, para cada medida especifico la siguiente informació: la dimensión a la que pertenece (`DenLex`, `SofLex` o `VarLex`), un identificador (Id.) de la métrica (primera columna), el nombre de la métrica (segunda columna), una etiqueta que facilita su posterior identificación en la sección de resultados (tercera columna), la fórmula para calcularla (cuarta columna) y alguna referencia bibliográfica (quinta columna). En la parte inferior del cuadro aparecen las convenciones necesarias para interpretar correctamente las fórmulas.

En la definición de las métricas asumo que las palabras de contenido léxico son los nombres, los adjetivos, los verbos y los adverbios. Dentro de las palabras de contenido gramatical estarían las preposiciones, las conjunciones, el artículo, los pronombres y los determinantes. De la misma forma, para determinar los índices de sofisticación (SofLex) me apoyo en una lista de las 5.000 formas más frecuentes del español de acuerdo con la RAE[19]. Según Laufer y Nation [1995], las palabras sofisticadas no están en la lista de las 1.000 o 2.000 palabras más frecuentes de un idioma. En este caso he usado las 5000 ya que la lista de la RAE contiene tanto palabras léxicas como gramaticales[20].

La mayoría de las métricas son transformaciones algebraicas que trabajan con índices léxicos para compensar la variación en la longitud de los textos. He optado por este método ya que considero que representa un nivel más avanzado en el cálculo de la riqueza léxica. Además, con ello evito tener que recurrir a métodos más simples de estandarización, por ejemplo, dividir todos los textos en segmentos de la misma longitud [Jarvis, 2002; Thordardottir y Weismer, 2001] o hacer una selección aleatoria de palabras [Breeder y col., 1986]. Estos últimos se consideran procedimientos «derrochadores» (*wasteful*) pues prescinden de datos potencialmente relevantes al tiempo que dificultan la reproducción de los experimentos y el contraste de los resultados [Lu, 2011; Malvern y col., 2004].

4.3.2 Análisis intra-textual

El análisis intra-textual consiste en establecer si existen diferencias léxicas significativas entre los comentarios que impidan asignarles un mismo registro lingüístico. En este análisis aplico las métricas de la riqueza léxica (`RiqLex`) descritas en el Apartado 4.3.1 sobre 1911 comentarios en español agrupados según tres criterios objetivos: el sexo, la edad y la procedencia de sus autores. Mi objetivo es determinar si alguna de estas tres variables demográficas permite identificar diferencias significativas a nivel del registro

19 `http://corpus.rae.es/lfrecuencias.html`

20 De ahí que todas las métricas de sofisticación léxica incorporen el «5» como parte de su etiqueta identificativa: SL5, PLF5, etc.

Id	Métrica	Etiqueta	Fórmula	Referencia
DenLex				
1	Densidad léxica	DL	$\frac{N_{lex}}{N}$	Engber [1995]
SofLex				
2	Sofisticacion léxica	SL5	$\frac{N_{slex}}{N_{lex}}$	Hyltenstam [1988]; Linnarud [1986]
3	Perfíl de Frecuencia Léxica	PFL5	$\frac{T_s}{T}$	Laufer y Nation [1995]
4	Sofisticación Verbal I	SV-I5	$\frac{T_{vs}}{N_v}$	Harley y King [1989]
5	Sofisticación Verbal II	SV-II5	$\frac{T_{vs}^2}{N_v}$	Chaudron y Parker [1990]
6	Sofisticación Verbal Corregida	SVC5	$\frac{T_{vs}}{\sqrt{2N_v}}$	Wolfe-Quintero y col. [1998]
VarLex				
7	Type/Token ratio	TTR	$\frac{T}{N}$	Templin [1957]
8	Root TTR	RTTR	$\frac{T}{\sqrt{N}}$	Guiraud [1960]
9	TTR Bilogarítmico	TTRB	$\frac{LogT}{LogN}$	Herdan [1960]
10	TTR Corregido	TTRC	$\frac{T}{\sqrt{2N}}$	J. Carroll [1964]
11	a^2	ac	$\frac{logN-logT}{log^2N}$	H. Maas [1972]; Tweedie y Baayen [1998]
12	Indice de Uber I	UI	$\frac{(logN)^2}{logN-logT}$	Dugast [1979]; Tweedie y Baayen [1998]
13	K de Yule	YuleK	$10^4 \times (\sum i^2 T_i - N_{lex})/N_{lex}^2$	Smith y Kelly [2002]; Tweedie y Baayen [1998]; Yule [1944]
14	Z de Zipf	ZIPF	$\frac{Z \times N \times log(N/Z)}{(N-Z)log(p \times Z)}$	Smith y Kelly [2002]; Tweedie y Baayen [1998]
15	Variación de Palabras Léxicas	VPL	$T_{lex} N_{lex}$	Engber [1995]
$16_{n,v,a,r}$	Variación G I	VG-I	$\frac{T_G}{N_{lex}}$	Harley y King [1989]; McClure [1991]
$20_{n,v,a,r}$	Variación G II	VG-II	$\frac{T_G}{N_G}$	Harley y King [1989]; McClure [1991]
24	Variación Mod.	VM	$\frac{(T_a+T_r)}{N_{lex}}$	Harley y King [1989]; McClure [1991]
$25_{n,v,a,r}$	Variación G Cuadrada	VG2-II	$\frac{T_G^2}{N_G}$	Wolfe-Quintero y col. [1998]
$29_{n,v,a,r}$	Variación G Corregida	VGC-II	$\frac{T_G}{\sqrt{2N_G}}$	Wolfe-Quintero y col. [1998]

N = tokens | T = types | $_{lex}$ = unidades léxicas | $_s$ = unidades sofisticadas

$_G$ = categoría gramatical (n, v, a, r) | $_n$ = nombre | $_v$ = verbo | $_a$ = adjetivo

$_r$ = adverbio | T_i = número de Types léxicos que ocurren i veces | Z = una medida de la riqueza léxica | p = Token más frecuente dividido por la longitud del texto

Cuadro 20: Métricas empleadas para el cálculo de la riqueza léxica.

lingüístico entre los comentarios. De no ser así, se infiere que esta clase de textos de opinión comparten el mismo registro lingüístico.

Los comentarios que analizo pertenecen al corpus HOpinion. Para asignar las etiquetas (demográficas) «edad», «sexo» y «procedencia» a cada texto, uso los metadatos del perfil público de cada usuario en TripAdvisor. En los pocos casos en los que el usuario no ha declarado uno de estos tres atributos demográficos, recurro a la información contextual para deducirlo. Por ejemplo, el alias «ANA1983Madrid» sugiere que se trata de una mujer de 32 años que vive en Madrid. En el Cuadro 21 presento la distribución de los datos atendiendo a la categoría (columna 1), las clases binarias[21] en las que cada categoría ha sido agrupada (columna 2), el número de muestras que han sido etiquetadas bajo cada clase (columna 3) y el número total de muestras por categoría (columna 4). Con el propósito de contar con una distribución alternativa de los datos, para edad he agrupado las muestras en dos subconjuntos. Un primer subconjunto, edad-1, con las clases $\geqslant$35 (1044 textos) y $\leqslant$34 (867 textos) y un segundo subconjunto (de control), edad-2, con las clases $\geqslant$50 (199 textos) y $\leqslant$24 (86 textos). Para propiciar una mayor separación o discretización de los datos, de este último subconjunto descarto 1626 muestras que corresponden a la franja de edad comprendida entre los 25 y los 49 años[22].

Categorías	Clases	Muestras	Total
Sexo	Hombres	948	1911
	Mujeres	963	
Edad-1	$\geqslant$35	1044	1911
	$\leqslant$34	867	
Edad-2	$\geqslant$50	199	285
	$\leqslant$24	86	
Procedencia	España	1450	1911
	América	461	

Cuadro 21: Distribución de las muestras usadas para el análisis intra-textual.

En los experimentos aplico técnicas de aprendizaje supervisado. La tarea consiste en predecir si un comentario x pertenece a alguna de las siguientes clases binarias: hombre/mujer, mayor_de_35/menor_de_34 años ($\geqslant$35 – $\leqslant$34), mayor_de_50/menor_de_24 años ($\geqslant$50 – $\leqslant$24) y español/latinoamericano (España – América). Para efectuar la clasificación me baso en una serie de atributos con los que han sido anotados los textos. En total realizo tres experimentos. En el primer experimento empleo las 32 medidas de riqueza léxica, de forma individual, como atributos para entrenar diferentes clasificadores. En el segundo experimento, agrupo las 32 medidas en cuatro dimensiones (`RiqLex`, `VarLex`, `SofLex` y `DenLex`) que uso como atributos de entrenamiento. Por último, en el tercer experimento, recurro a la selección de rasgos con el propósito de identificar las métricas más adecuadas para clasificar los comentarios. En los experimentos uso Weka [Witten y col., 1999]. Las pruebas se confeccionaron escogiendo 90 % de datos para el entrenamiento (*train*) y 10 % para validar las pruebas (*test*). Todos los clasificadores fueron

21 Por ejemplo, la categoría edad, puede agruparse en dos clases: menores de 34 versus mayores de 35 años.

22 A diferencia de las categorías sexo y origen, la edad es una categoría que asume valores continuos y con Edad-2 estoy forzando la discretización de los datos.

entrenados usando validación cruzada (*ten-fold cross-validation*). El resultado de los clasificadores se da en términos de su precisión (*Prediction Accuracy*).

4.3.2.1 *Experimento 1*

En este experimento efectúo un análisis aplicando las 32 métricas de `RiqLex` (ver Cuadro 20) directamente sobre los textos agrupados por categorías (sexo, procedencia y edad) y clases (hombre/mujer, España/América, etc.). Los valores que retorna cada una de las métricas los utilizo como conocimiento para entrenar una serie de clasificadores (ver Cuadro 22).

Algoritmos de aprendizaje
1. Bayes: BayesNet, BayesianLogisticRegression, NaiveBayes, NaiveBayesSimple, NaiveBayesUpdateable
2. Perezosos: IB1, IBk, KStar, LWL
3. Reglas: ConjunctiveRule, DTNB, DecisionTable, JRip, NNge, OneR, PART, Ridor, ZeroR
4. Árboles: ADTree, BFTree, DecisionStump, FT, J48, J48graft, LADTree, LMT, NBTree, REPTree, RandomForest, RandomTree, SimpleCart
5. Misceláneos: HyperPipes, VFI

Cuadro 22: Lista de los algoritmos de clasificación usados (33 en total).

En el Cuadro 23 presento los resultados de este primer experimento. Cada celda contiene la precisión de los 10 *folds* (*Prediction Accuracy*, PA) obtenida al clasificar los comentarios en las clases binarias asociadas a cada uno de las categorías: sexo (hombre/mujer), edad-1 ($\geqslant$35/$\leqslant$34), edad-2 ($\leqslant$24/$\geqslant$50) y procedencia (español/latinoamericano). En la primera fila tenemos el PA de los mejores clasificadores, en la segunda fila el PA de los peores y en la última fila el PA promedio de los 330 clasificadores obtenidos mediante *10-fold cross-validation* para cada categoría.

	sexo	**edad-1**	**edad-2**	**procedencia**
mejor	60 %	57.3 %	98.8 %	77.3 %
peor	53.1 %	52.6 %	67 %	58.4 %
promedio	55.6 %	54.9 %	81 %	73.7 %

Cuadro 23: Precisión (PA) obtenida en el primer experimento.

Como se puede ver en el Cuadro 23, en general, no existen diferencias significativas a nivel de registro entre los comentarios después de agruparlos en las diferentes variables demográficas. El rendimiento promedio de los clasificadores para las categorías sexo y edad-1 no supera el 60 %: 55.6 % y 54.9 % respectivamente, lo cual indica que los comentarios no se pueden discriminar mediante el uso de estas dos categorías. La única categoría que ha mostrado cierta capacidad discriminatoria es la procedencia de los autores de los comentarios. Efectivamente, el rendimiento de los clasificadores mejora cuando la categoría a predecir es procedencia, su PA promedio es del 73.7 %. No obstante, aún seguimos encontrando valores inferiores al 60 % en alguno de los *folds*. Por último, la categoría edad-2 se ha de analizar aparte. Esta categoría ha obtenido un rendimiento promedio del 81 % gracias a que, como he comentado anteriormente, se ha forzado la agrupación de los datos

en dos conjuntos bien disjuntos: mayores de 50 versus menores de 24 años. El objetivo de edad-2 es usarlo como un grupo de control.

4.3.2.2 *Experimento 2*

Empleando los mismos algoritmos de clasificación del experimento anterior (ver Cuadro 22), en este experimento entreno de nuevo los clasificadores agrupando las métricas según la dimensión de la `RiqLex` que evalúan. El objetivo es determinar si alguna dimensión de la `RiqLex` puede ayudar a clasificar de manera más precisa los comentarios según las categorías demográficas seleccionadas.

En este experimento obtengo tres grupos de clasificadores, uno por cada dimensión. El primer grupo aprende de una única métrica de `DenLex` (DL) y de dos de las categorías demográficas (edad-1/edad-2, sexo o procedencia) que incorporo como parte de su conocimiento. El segundo grupo aprende de las métricas de `SofLex` (de la 2 a la 6 en el Cuadro 20) y, también, de dos de las categorías demográficas. El tercer grupo extrae el conocimiento de las medidas de `VarLex` (de la 7 a la 29) y, nuevamente, de dos categorías demográficas. En todos estos casos, la categoría demográfica omitida es el atributo a aprender o atributo de clase.

En el Cuadro 24 expongo los promedios de predicción obtenidos con todos los clasificadores entrenados usando validación cruzada de diez iteraciones. Los promedios aparecen agrupando según las clases a aprender (filas) y las dimensiones de la riqueza léxica evaluadas (columnas). En la primera columna del cuadro (`RiqLex`) se mantienen los promedios de predicción obtenidos en el experimento anterior y que resultan de aplicar todas las métricas simultáneamente (ver Cuadro 23). Los números entre paréntesis señalan los rangos (R), es decir, la bondad de la predicción de cada dimensión para cada una de las categorías evaluadas (sexo, edad-1, edad-2 y procedencia): el rango (1) indica el mejor PA y el rango (4) el peor en esa fila. La suma de los rangos, en la última fila (Rango), refleja la posición definitiva que ocupa cada dimensión: a menor rango, mejor predicción.

	RiqLex		**VarLex**		**SofLex**		**DenLex**	
	PA	R	PA	R	PA	R	PA	R
sexo	55.6	(1)	54.8	(2)	54.4	(3)	53.7	(4)
edad-1	54.9	(2)	53.6	(4)	53.9	(3)	56.4	(1)
edad-2	81	(1)	80.2	(2)	78.3	(3)	76.9	(4)
procedencia	73.7	(3)	74.6	(2)	73.7	(4)	74.7	(1)
Rango		**(7)**		**(10)**		**(13)**		**(10)**

Cuadro 24: Precisión (PA) obtenida para cada categoría en el análisis inter-textual.

Como se observa en el Cuadro 24, la `RiqLex` obtiene los mejores rangos ((1), (2), (1), (3)) que cualquiera de las dimensiones que la componen de forma individual. Esto significa que las 32 métricas en conjunto funcionan mejor que por separado. No obstante, de cara a reducir aún más el número de atributos con los que se trabaja, las métricas de `VarLex` o de `DenLex` predicen mejor que las de `SofLex`. En efecto, si atendemos a los rangos de la `SofLex` ((3), (3), (3), (4)) comprobaremos que las cinco métricas que la componen son las menos efectivas para clasificar los comentarios.

Los resultados de este análisis indican que la `RiqLex` (suma de todas las dimensiones) es la opción más efectiva para medir la variación léxica en-

tre comentarios. `VarLex` y `DenLex` estarían en un segundo plano, con niveles similares de rendimiento. Finalmente, la sofisticación léxica (`SofLex`) es la métrica menos efectiva para clasificar los comentarios según la edad, el sexo y la procedencia de sus autores. Por lo tanto, como sucedió en el experimento 1, los PAs obtenidos no revelan diferencias léxicas sustanciales entre los comentarios, a excepción de la categoría demográfica edad-2 que, como ya he comentado, agrupa los textos en clases bien diferenciadas.

4.3.2.3 *Experimento 3*

En el tercer experimento recurro a la técnica de selección de rasgos con el objetivo de detectar las métricas más adecuadas para clasificar los comentarios según las categorías demográficas. Por este motivo, en este caso selecciono solo las categorías que obtuvieron los mejores niveles de precisión en los dos experimentos anteriores, esto es edad 2 y procedencia.

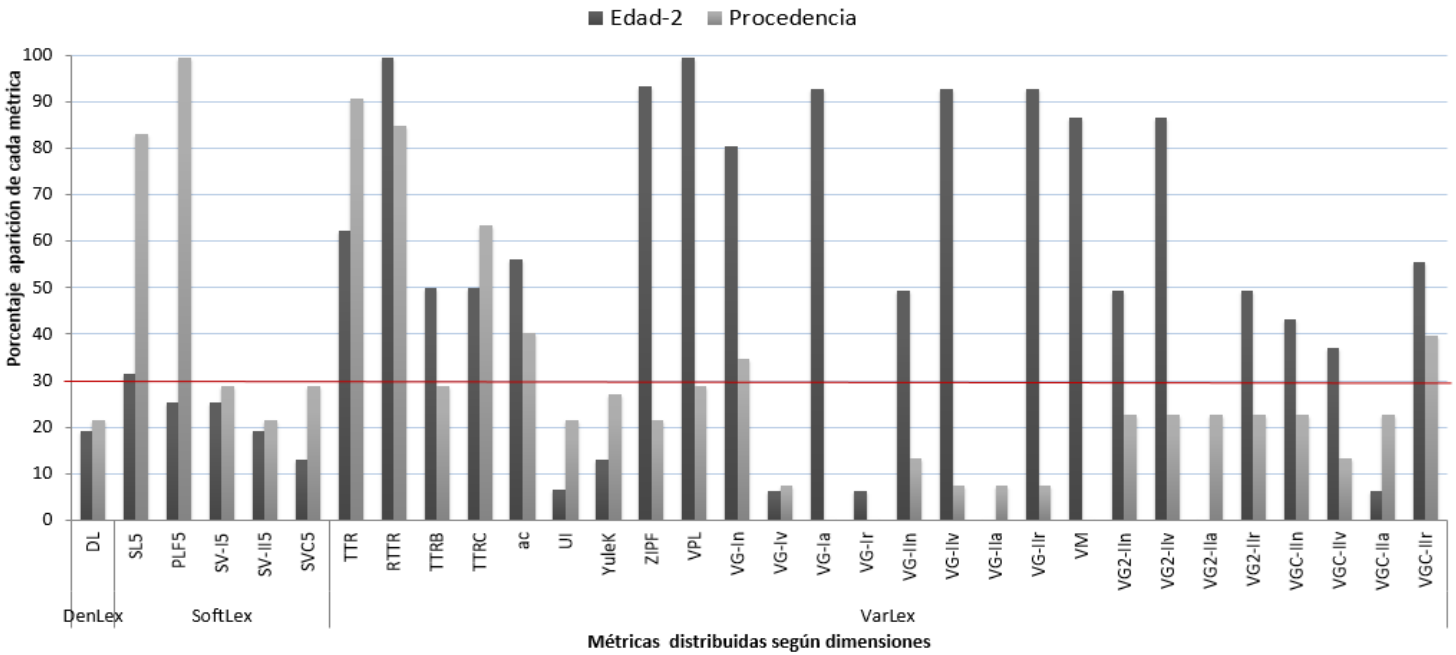
La configuración de este experimento es la siguiente. Empleo los 13 mejores algoritmos de clasificación de los experimentos anteriores, conjuntamente con 7 métodos evaluadores y 5 selectores (ver la nota al pie en la Figura 7). En el caso de Ranker, el selector se configuró para que recuperara solo los 25 atributos/métricas más discriminantes de las 32 existentes ya que por defecto solo pondera y se ha de establecer manualmente el nivel de selección.

En primer lugar, en la Figura 7 aparece el porcentaje de uso que los mejores clasificadores hacen de cada una de las métricas agrupadas según las tres dimensiones de la `RiqLex`: `DenLex`, `SofLex` y `VarLex`. Por ejemplo, se observa que 20 de las 32 métricas están por encima de un 30 % de aparición para edad-2 (SL5, TTR, RTTR, TTRB, TTRC, ac, etc.), mientras que solo 8 de las métricas están por encima de este porcentaje para procedencia (SL5, PLF5, TTR, RTTR, etc.). Esto significa que hace falta más información para clasificar los comentarios por la categoría edad-2 que por procedencia. Otras métricas como VG-Ia solo la utilizan los clasificadores de edad-2.

En segundo lugar, en la misma Figura 7 vemos que para predecir edad-2 son más útiles las métricas de la `VarLex` basadas en categorías gramaticales como VG-Ir, VG-IIa y VG2-IIa. Por su parte, las métricas más frecuentes para predecir la procedencia corresponden al TTR tradicional y sus variaciones (RTTR, TTRB y TTRC) así como a dos métricas de la `SofLex` (SL5 y PLF5).

De acuerdo con estos resultados podemos afirmar que el análisis de las palabras con contenido léxico es la opción más adecuada para clasificar usuarios por edad dada la alta frecuencia de aparición de las medidas que utilizan este tipo de palabras. Además, los resultados apuntan a que una clasificación basada en el análisis de las frecuencias de palabras léxicas conjuntamente con las gramaticales favorece, hasta cierto punto, la detección de usuarios por procedencia.

La conclusión general que se extrae de estos tres experimentos que forman parte del análisis intra-textual de los comentarios es que no existen diferencias léxicas significativas asociadas con el sexo, la edad o la procedencia de los autores de estos textos, lo cual es un fuerte indicador de que no presentan diferencias estilísticas destacables. Si bien, mediante el análisis intra-textual he probado múltiples formas de agrupar los comentarios, siempre con los mismos resultados, el análisis inter-textual me permitirá contrastar el registro de los comentarios con el de otro tipo de textos.



(a) DMNBtext, NaiveBayes, IBk, rules.DTNB / OneR / PART, trees.ADTree / FT / J48 / J48graft / LADTree / LMT / RandomForest. (b) Cfs, ChiSquared, Consistency, Filtered, InfoGain, PCA, ReliefF. (c) BestF., GeneticS., LinearForwardS., RankS., Ranker.

Figura 7: Porcentaje de uso de las métricas según los mejores clasificadores. Adicionalmente, en (a) lista de algoritmos, en (b) lista de evaluadores, y en (c) lista de selectores usados en los experimentos.

4.3.3 Análisis inter-textual

El análisis inter-textual consiste en establecer si existen diferencias léxicas significativas entre los comentarios y otra clase de textos que impidan asignarles un mismo registro lingüístico. En este análisis selecciono un conjunto de las métricas de riqueza léxica descritas en el Apartado 4.3.1 que aplico sobre 1635 comentarios y 1635 artículos periodísticos. Mi objetivo es determinar si es posible identificar diferencias significativas a nivel del registro lingüístico entre los comentarios y los textos periodísticos.

Los comentarios que analizo pertenecen al corpus HOpinion y los artículos periodísticos al corpus AnCora-ES. *A priori*, estos dos corpus representan registros diferentes del español actual: HOpinion, registro coloquial y AnCora-ES, registro formal. El Cuadro 25 describe las características principales del conjunto de datos utilizado para este análisis. Para una descripción más detallada de ambos corpus ver las secciones 3.2.1 y 3.3.2.1.

Característica	HOpinion	AnCora-Es
Registro	Coloquial	Formal
Género	Comentario	Artículo
Fuente de datos	TripAdvisor	El Periódico & Agencia EFE
Número de textos	1635	1635
Total palabras	206.812	443.380
Palabras por texto (aprox.)	126	271
Idioma	Español	Español

Cuadro 25: Descripción del conjunto de datos utilizado para el análisis inter-textual.

En los experimentos aplico técnicas de aprendizaje supervisado. La tarea consiste en predecir si un texto x pertenece a la clase «comentario» (textos *a priori* coloquiales) o «artículo» (textos *a priori* formales) basándose en una serie de atributos con los que han sido anotados los textos. En total realizo dos experimentos. En el primer experimento selecciono algunas de las métricas más representativas de cada una de las tres dimensiones que componen la riqueza léxica y las empleo como atributos para entrenar diferentes clasificadores. En el segundo experimento evalúo una aproximación alternativa para identificar posibles diferencias a nivel del registro lingüístico entre los comentarios y los textos periodísticos. En este segundo experimento combino información léxica con información morfosintáctica para entrenar los clasificadores. En ambos experimentos, como herramienta de clasificación, uso Weka [Witten y col., 1999]. Concretamente, selecciono 38 conocidos algoritmos de aprendizaje, los cuales se pueden dividir en 5 grandes categorías (ver Cuadro 26). La evaluación la realizo mediante la técnica de validación cruzada de 10 iteraciones (*10-fold cross-validation*). Las pruebas se confeccionaron escogiendo 90 % de datos para el entrenamiento (*train*) y 10 % para validar las pruebas (*test*). El resultado de los clasificadores se da en términos de su precisión (*Prediction Accuracy*), es decir, el porcentaje de instancias que fueron correctamente clasificadas como comentario (registro coloquial) o artículo periodístico (registro formal). Con el fin de determinar los atributos que tienen más peso, aplico varios métodos de selección de atributos. Los métodos de selección de atributos reducen el número de variables, seleccionando el mejor subconjunto de características del conjunto inicial. Los métodos de evaluación y de búsqueda usados en la selección supervisada se enumeran, también, en el Cuadro 26.

Algoritmos de Aprendizaje
1. Bayes: BayesNet, BayesianLogisticRegression, ComplementNaiveBayes, DMNBtext, NaiveBayes, NaiveBayesMultinomial, NaiveBayesMultinomialUpdateable, NaiveBayesSimple, NaiveBayesUpdateable **2. Perezosos:** IB1, IBk, KStar, LWL **3. Reglas:** ConjunctiveRule, DTNB, DecisionTable, JRip, NNge, OneR, PART, Ridor, ZeroR **4. Árboles:** ADTree, BFTree, DecisionStump, FT, J48, J48graft, LADTree, LMT, NBTree, REPTree, RandomForest, RandomTree, SimpleCart, lmt.LogisticBase **5. Misceláneos:** HyperPipes, VFI
Métodos de evaluación
CfsSubsetEval, ChiSquaredAttributeEval, ConsistencySubsetEval, FilteredAttributeEval, FilteredSubsetEval, GainRatioAttributeEval, InfoGainAttributeEval, LatentSemanticAnalysis, OneRAttributeEval, PrincipalComponents, ReliefFAttributeEval, SVMAttributeEval, WrapperSubsetEval
Métodos de búsqueda
BestFirst, ExhaustiveSearch, GeneticSearch, GreedyStepwise, LinearForwardSelection, RandomSearch, RankSearch, Ranker, ScatterSearchV1, SubsetSizeForwardSelection

Cuadro 26: Listado de los algoritmos, métodos de evaluación y de búsqueda utilizados para la clasificación del registro lingüístico.

4.3.3.1 *Experimento 1*

En este experimento efectúo un análisis de la riqueza léxica sobre los textos agrupados según el corpus del que proceden: HOpinion y AnCora-ES. De las 32 métricas de la riqueza léxica presentadas en el Apartado 4.3.1 (ver Cuadro 20) he seleccionado nueve: la de densidad léxica (`DenLex`), dos de sofisticación léxica (SL5 y PFL5) y seis de variación léxica (TTR, ac, UI, ZIPF, VPL y VM). He seleccionado estas métricas, representativas de cada dimensión, puesto que dieron los mejores resultados para clasificar los comentarios según los atributos demográficos[23] (análisis intra-textual). Para calcular las métricas he usado la herramienta AToP[24] [Singateh, 2013][25], desarrollada como parte de esta investigación.

Adicionalmente, teniendo en consideración que los comentarios y los artículos periodísticos se suelen asociar con registros lingüísticos contrapuestos, he agregado al conjunto de rasgos dos características más que tienen que ver con la detección de términos coloquiales y términos de emoción. Los términos con un uso coloquial los obtuve consultando las versiones en línea de Wikcionario[26] y TheFreeDictionary[27]. Las marcas de uso que hago servir para reconocer dichos términos en ambos diccionarios son: «coloquial», «despectivo», «malsonante», «familiar», «informal», «peyorativo» y «vulgar». La detección de los términos de emoción la efectué mediante el *Spanish Emotion Lexicon* [Sidorov y col., 2012], un recurso léxico creado de forma totalmente manual por investigadores del Instituto Politécnico Nacional de México. En la Figura 8 presento dos ejemplos del formato (XML) y el tipo de in-

23 La `DenLex` no obtuvo buenos resultados en el análisis intra-textual pero la incluimos en el análisis por coherencia, pues es la única métrica de la que disponemos para evaluar esta dimensión de la riqueza léxica.

24 Para mayores detalles sobre el diseño y configuración de AToP (Análisis de Textos de Opinión en lenguaje natural), recomiendo la lectura del Anexo A.

25 Este documento hace referencia a un proyecto desarrollado en la Universidad de Barcelona, basado en mi trabajo y en el que participé como asesor.

26 `http://es.wiktionary.org/`

27 `http://es.thefreedictionary.com/`

formación con que se ha anotado cada una de las palabras en los corpus AnCora-Es y HOpinion a partir de los recursos señalados. En la parte superior del cuadro aparecen las palabras «acojonante» y «cabrear» asociadas con su Factor de Probabilidad de uso Afectivo (*Probability Factor of Affective use*, PFA) según *Spanish Emotion Lexicon*. En la parte intermedia del cuadro tenemos las entradas correspondientes en los diccionarios *online* TheFreeDictionary y Wikcionario. En la parte baja del cuadro tenemos el resultado final del proceso automático de anotación.

Palabra	**Nula[%]**	**Baja[%]**	**Media[%]**	**Alta[%]**	**PFA**	**Categoría**
acojonante	10	0	50	40	0.73	Sorpresa
cabrear	20	10	40	30	0.597	Enojo

THE FREE DICTIONARY BY FARLEX	Wiktionary The free dictionary
acojonante: *adj. vulg.* Que impresiona mucho, positiva o negativamente.	**cabrear:** Enfadar o poner de mal humor. - Ámbito: España. - Uso: coloquial, malsonante, se emplea también como pronominal: cabrearse.

```
<wd
  lemma="acojonante"
  register="coloquial"
  emotion="sorpresa"
  source="hopinion"
/>
```

```
<wd
  lemma="cabrear"
  register="coloquial"
  emotion="enojo"
  source="ancora"
/>
```

Figura 8: Ejemplos de palabras anotadas con registro y emoción.

En total, son once los atributos seleccionados en este experimento. En el Cuadro 27 reproduzco el listado completo de estos atributos que utilizo para el entrenamiento de diversos clasificadores. Los atributos o características están organizados según la dimensión de la riqueza léxica que evalúan (`DenLex`, `SofLex` y `VarLex`). Además, se incorpora el otro par de atributos seleccionados para este experimento, los términos de `registro` y de `emoción`. En la primera columna del cuadro aparecen las dimensiones evaluadas, en la segunda un identificador del atributo (Id), en la tercera columna el nombre de la métrica (o la característica) y en la cuarta columna la etiqueta asociada a cada atributo. Por último, la quinta columna hace referencia a la fórmula que se usa para calcular cada métrica según lo expuesto en el Cuadro 20.

A continuación describo los niveles de precisión alcanzados al predecir el registro lingüístico de los textos en los corpus HOpinion y AnCora-Es, usando las once características enumeradas anteriormente. El total de clasificadores que entreno en este experimento es de 13.768. En general, los clasificadores que usan algún método de selección de atributos y los que no los usan se comportan de manera similar. En ambos casos la precisión promedio es del 84 % (ver Figura 9).

En el Cuadro 28 relaciono la precisión con el número de atributos seleccionados. En el cuadro aparecen los algoritmos, métodos de selección y búsqueda, que permitieron obtener la mejor precisión (PA) de acuerdo con el

Dimensión	Id	Característica	Etiqueta	Ref.
DenLex	1	Densidad Léxica	DL	1
SofLex	2	Sofisticación Léxica	SL5	2
	3	Perfil de Frecuencia Léxica	PFL5	3
VarLex	4	Type Token Ratio	TTR	7
	5	α^2	ac	11
	6	Índice de Uber	IU	12
	7	Z de Zipf	ZIPF	14
	8	Variación de palabras léxicas	VPL	15
	9	Variación modal	VM	24
Otras	10	Registro	USO	–
	11	Emoción	EMO	–

Cuadro 27: Características evaluadas en los textos de los corpus HOpinion y AnCora-ES para el análisis inter-textual (experimento 1).

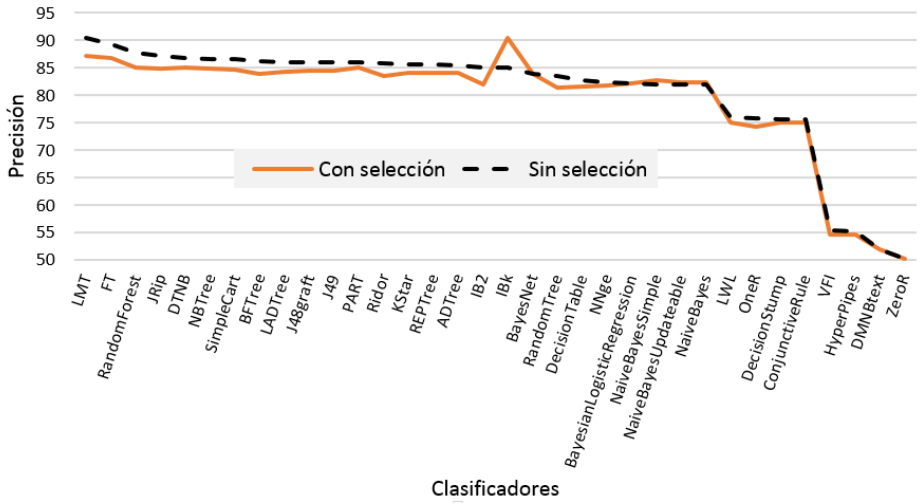


Figura 9: Precisión con y sin selección de atributos (promedios).

número de atributos (#att) y el rendimiento promedio de todos los clasificadores que usan un método de selección de atributos ($\overline{\text{PA}}$). Como se puede observar todos los clasificadores, excepto los que usan un sólo atributo, superan el 80 % de precisión. Los PAs más altos, también de acuerdo con el número de atributos, superan el 90 % y usan algoritmos basados en árboles de clasificación (*trees*).

Algoritmo	Mét. evaluación	Mét. búsqueda	PA+	$\overline{\text{PA}}$	#att
trees.FT	ChiSquaredAttributeEval	Ranker	93.5	80.6	11
trees.RandomForest	ConsistencySubsetEval	RankSearch	93.8	80.0	10
trees.LMT	ConsistencySubsetEval	RandomSearch	93.5	81.5	9
trees.FT	PrincipalComponents	Ranker	94.1	80.7	7
trees.LMT	CfsSubsetEval	GeneticSearch	92.0	81.2	6
trees.FT	CfsSubsetEval	BestFirst	92.6	80.1	5
lazy.KStar	WrapperSubsetEval	RankSearch	79.1	64.6	1

Cuadro 28: Rendimiento de los atributos léxicos.

Las figuras 10 a la 12 sintetizan, mediante tres diagramas de caja (*box-plots*), los principales resultados de este experimento. Debido a que de los 11 atributos léxicos (ver Cuadro 27) nueve tiene que ver con la riqueza léxica (RIQ), uno con el uso de términos característicos del registro coloquial (USO) y uno con el uso de los términos de emoción (EMO), en la Figura 10 presento la relación que hay entre estas tres clases de atributos y la precisión. Como se puede observar, la mediana más alta la tienen los clasificadores que usan las tres clases de atributos (RIQ+USO+EMO), luego están los que utilizan RIQ conjuntamente con USO (RIQ+USO), seguidos por los clasificadores que utilizan solo EMO o RIQ. Ningún clasificador emplea el atributo USO de forma aislada.

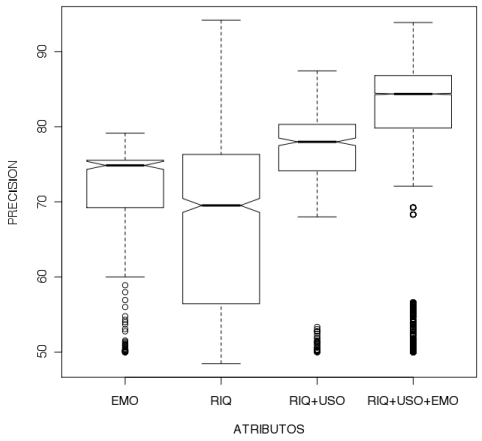


Figura 10: Precisión en relación con los atributos «emoción» (EMO), «riqueza» (RIQ) y «uso» (USO).

En la Figura 11 presento los atributos que han seleccionado los 13.768 clasificadores como los más relevantes. En total son seis atributos: el de emoción (EMO), el del uso coloquial (USO) y cuatro de riqueza léxica: Densidad Léxica (DL), Sofisticación Léxica (SL5), Perfil de Frecuencia Léxica (PFL5) y Variación Modal (VM). En este último caso la Densidad Léxica (DL), junto con el uso coloquial (USO), ha dado mejores resultados.

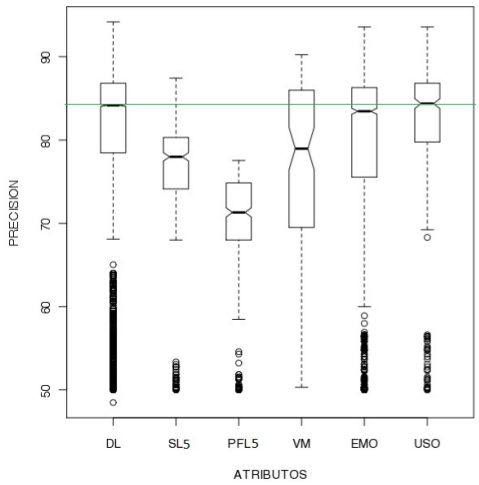


Figura 11: Mejores atributos léxicos seleccionados por los clasificadores.

En la Figura 12 relaciono la precisión con el número de atributos seleccionados. Los clasificadores que tienen un mejor rendimiento usan nueve, seis y siete atributos. Estos clasificadores presentan precisiones promedio del 85 %.

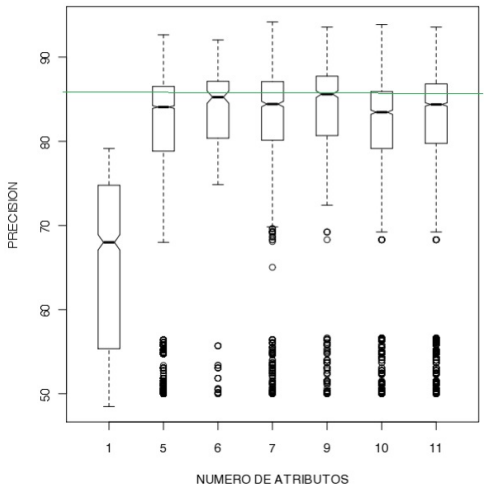


Figura 12: Precisión en relación con el número de atributos léxicos utilizados por los clasificadores.

4.3.3.2 *Experimento 2*

En este experimento aplico una serie de características morfosintácticas, conjuntamente con dos rasgos de riqueza léxica, para comparar el registro lingüístico de los comentarios con el de los artículos periodísticos. Las características morfosintácticas seleccionadas representan seis de los fenómenos lingüísticos más documentados en los estudios sobre el español coloquial. A continuación defino cada uno de estos fenómenos y en el Cuadro 29 enumero las características que se derivan de tales fenómenos.

1. Sintaxis concatenada (SXC): consiste en la acumulación de enunciados producto de la ausencia de planificación en la producción del mensaje [Narbona, 1989]. La ausencia de signos de puntuación y el predominio de la coordinación frente a la subordinación (sintaxis incrustada), son dos de las consecuencias más evidentes de la sintaxis concatenada.

2. Elipsis (ELP): consiste en la omisión de elementos lingüísticos que se presuponen a partir de entidades que se hallan presentes en el contexto discursivo. Una forma básica pero muy frecuente de representar tales omisiones es mediante el uso de los puntos suspensivos. Las oraciones suspendidas, tan características del lenguaje coloquial, son una claro ejemplo de ello: «si yo les contara...», «vamos, que en vuestro lugar yo no....».

3. Redundancia (RED): consiste en repetir, de manera exacta o aproximada, algunos elementos, desde vocales o consonantes («nooooo»), pasando por las palabras («es un hostal cutre, cutre, cutre») hasta llegar a los sintagmas («son unos guarros unos guarros») [Tannen, 1989].

4. Deixis (DXS): es un recurso para la cohesión textual que el hablante utiliza para introducir las entidades o referentes del contexto situacional

en el discurso. Esta función es expresada por pronombres personales y demostrativos. Por ejemplo, el pronombre sujeto de primera persona suele aparecer en textos subjetivos y coloquiales conjuntamente con verbos de opinión («yo opino», «yo creo»).

5. Riqueza léxica (`RiqLex`): son diferentes métricas que se utilizan para conocer la competencia léxica de un hablante [Read, 2005]. En general, se considera que la riqueza léxica de los textos coloquiales es baja, si se les compara con textos formales como los científicos o literarios.

6. Intensificación (INT): la expresión de emociones y sentimientos es típica del lenguaje coloquial. La expresión de emociones suele asociarse al uso abundante de interjecciones («¡buah...!», «¡ufff!») y de oraciones consecutivas intensivas («es una habitación tan grande que...»).

Las características que se derivan de estos seis fenómenos lingüísticos y que usaré como atributos para el entrenamiento de diversos clasificadores son las que aparecen en el Cuadro 29. Para calcular las frecuencias de estas once características[28], los textos se han etiquetado con Part-Of-Speech (POS) y lema. La riqueza léxica (`RiqLex`) la obtengo mediante el cálculo de la densidad léxica (DL) y la variación léxica[29]. Como en los otros experimentos de este capítulo, las fórmulas para calcular DL y TTR aparecen en el Cuadro 20 de la página 52.

N°	Característica	Fenómeno lingüístico
1	Densidad léxica (DL)	`RiqLex`
2	Signos de puntuación	SXC
3	Repetición de palabras y sintagmas	RED
4	Conjunciones coordinantes	SXC
5	Conjunciones subordinantes	SXC
6	Pronombres personales y demostrativos	DXS
7	Puntos suspensivos	ELP
8	Interjecciones	PAR
9	Repetición de vocales y consonantes	RED
10	Oraciones consecutivas intensivas*	INT
11	Variación léxica (TTR)	`RiqLex`

*Para identificar los esquemas oracionales consecutivos he utilizado el siguiente patrón: *[intensificador: tanto, tan, tal, etc.] + [nombre OR adjetivo OR adverbio] + [que]*

Cuadro 29: Características morfosintácticas evaluadas en los textos.

A continuación describo los niveles de precisión alcanzados al predecir el registro lingüístico de los textos en los corpus HOpinion y AnCora-Es, usando las once características enumeradas anteriormente. En primer lugar, aplicando la configuración detallada al principio de esta sección, entreno un total de 14.400 clasificadores[30]. En promedio, los clasificadores que usan algún método de selección de atributos funcionan mejor que los que no los usan (ver Figura 13).

28 En este análisis no empleamos ninguna medida de la sofisticación léxica puesto que esta dimensión de la riqueza léxica suele relacionarse con la adquisición de segundas lenguas antes que con el registro lingüístico.

29 Concretamente *type-token ratio*.

30 La combinación de algunos métodos de evaluación y de búsqueda no son posibles en Weka.

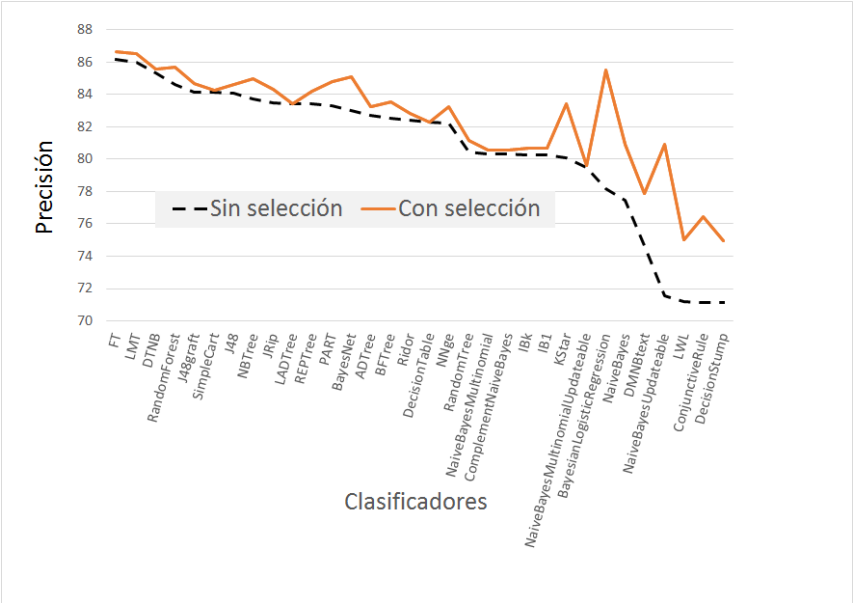


Figura 13: Precisión con y sin selección de atributos (promedios). Usando la selección de atributos se consigue reducir el número de características, mientras que la precisión mejora o se mantiene.

En segundo lugar, para identificar los atributos con mayor valor predictivo, en el Cuadro 30 contrasto el rendimiento promedio de los clasificadores cuando usan un determinado atributo (A_x) y cuando dejan de usarlo ($\neg A_x$). De esta manera, el valor Diferencia (Dif.) determinará el impacto que tiene la omisión del atributo x a nivel de la precisión. Los atributos más informativos son el 4 y el 6, es decir, conjunciones coordinantes y pronombres personales y demostrativos. Los menos informativos son el 2 y el 9 (signos de puntuación y repetición de vocales y consonantes).

		Promedios		
Id	Atributo	A_x	$\neg A_x$	Dif.
1	Densidad léxica (DL)	78.3 %	65.5 %	12.8
2	Signos de puntuación	78.5 %	76.1 %	2.4
3	Repetición de palabras y sintagmas	78.1 %	72.7 %	5.4
4	Conjunciones coordinantes	78.6 %	62.2 %	16.4•
5	Conjunciones subordinantes	78.6 %	76.3 %	2.3
6	Pronombres personales y demostrativos	78.6 %	62.2 %	16.4•
7	Puntos suspensivos	78.6 %	65.7 %	12.9
8	Interjecciones	78.5 %	71.3 %	7.2
9	Repetición de vocales y consonantes	78.6 %	76.2 %	2.4
10	Oraciones consecutivas intensivas	78.6 %	76.1 %	2.5
11	Variación léxica (TTR)	78.6 %	65.9 %	12.7

Cuadro 30: Atributos morfosintácticos con mayor valor predictivo.

Finalmente, para conocer el rendimiento de los clasificadores, en el Cuadro 31 relaciono la precisión con el número de atributos seleccionados. En el cuadro aparecen los algoritmos, métodos de selección y búsqueda, que permitieron obtener la mejor precisión (PA) de acuerdo con el número de atributos (#att) y el rendimiento promedio de los 14.400 clasificadores ($\overline{PA}$). Adicionalmente, en la Figura 14 se presenta la relación precisión - número

de atributos de forma gráfica. Como se observa tanto en el cuadro como en la figura, las configuraciones que usan 6, 7, 8, 10 y 11 atributos presentan niveles de precisión superiores al 85 %. Si bien, la precisión más alta se obtiene con 10 atributos (86.6 %) de los 11 disponibles, la mejor relación **precisión versus número de atributos** se logra usando solo 6 o 7 atributos. De manera específica, con seis atributos se ha podido conseguir una precisión del 85.9 % y con siete del 86 %, una diferencia mínima que favorece a la configuración con menos atributos.

Algoritmo	Mét. evaluación	Mét. búsqueda	PA+	$\overline{PA}$	#att
trees.FT	SVMAttributeEval	Ranker	86.1	78.6	11
trees.FT	ConsistencySubsetEval	SubsetSizeForwardSelection	86.6	78.6	10
trees.FT	CfsSubsetEval	BestFirst	85.6	78.6	8
trees.FT	FilteredSubsetEval	BestFirst	86	78.7	7
trees.LMT	FilteredSubsetEval	SubsetSizeForwardSelection	85.9	78.5	6
trees.FT	WrapperSubsetEval	RandomSearch	82.1	76.4	5
lazy.KStar	LatentSemanticAnalysis	Ranker	72	62.2	1

Cuadro 31: Rendimiento de los atributos morfosintácticos.

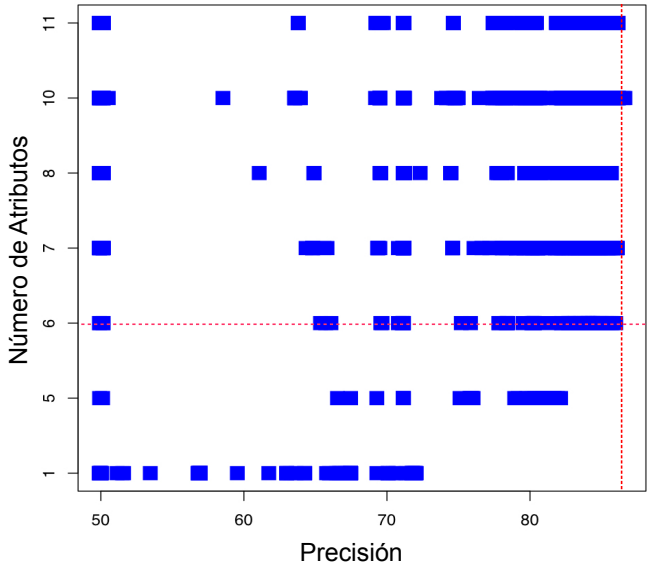


Figura 14: Rendimiento de los atributos morfosintácticos (representación gráfica).

La conclusión general que se extrae de estos dos experimentos que forman parte del análisis inter-textual de los comentarios es que existen diferencias léxicas y morfosintácticas significativas entre los comentarios y los artículos periodísticos, lo cual es un fuerte indicador de que ambos tipos de textos presentan diferencias estilísticas destacables. Como se observa en el diagrama de cajas (*box plot*) de la Figura 15, los resultados obtenidos tras el análisis inter-textual (comentarios versus artículos periodísticos) superan a los del análisis intra-textual (comentarios versus comentarios) demostrando que los comentarios son textos poco disociables entre sí y altamente disociables si se les compara con otra clase de textos. En el diagrama de cajas la mediana del

análisis inter-textual, es decir comentarios versus artículos periodísticos, es elevada (84 %) mientras que la del análisis intra-textual, comentarios versus comentarios, es significativamente más baja (67 %). De todo lo expuesto se puede concluir que los comentarios comparten el mismo registro lingüístico.

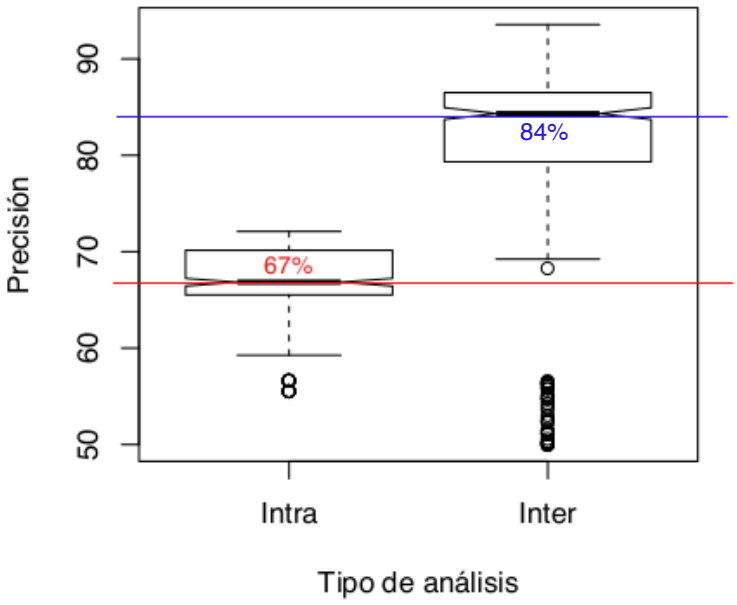


Figura 15: Resumen del análisis contrastivo del registro lingüístico: comentarios versus comentarios (análisis intra-textual) y comentarios versus artículos periodísticos (análisis inter-textual).

Para finalizar este capítulo, presento las conclusiones generales que se obtienen del análisis de la estructura discursiva y del registro lingüístico de los comentarios sobre productos que forman parte de su caracterización como un género discursivo.

4.4 CONCLUSIONES

En este capítulo he analizado la estructura discursiva y del registro lingüístico de los comentarios con el objetivo de determinar si este tipo de textos de opinión constituyen un género discursivo propio.

- Al analizar un conjunto representativo de comentarios, he constatado que presentan una estructura discursiva relativamente estable caracterizada por la presencia de tres tipos de segmentos: narrativo, descriptivo y valorativo.
- Al contrastar el registro lingüístico de un conjunto representativo comentarios entre sí (análisis intra-textual), no ha sido posible identificar diferencias léxicas destacables que indiquen que estamos ante el mismo tipo de texto. Por el contrario, al contrastar el registro lingüístico

de los comentarios con artículos periodísticos (análisis inter-textual), he constatado que existen diferencias léxicas e, incluso, diferencias morfosintácticas significativas que nos indican que estamos ante dos tipos diferentes de textos.

Estos resultados me permiten concluir que los comentarios sobre productos conforman un género discursivo propio. A nivel de estructura, el segmento valorativo caracteriza este género puesto que los comentarios tienen como propósito comunicativo principal valorar positiva o negativamente un producto. Dicho propósito comunicativo se consigue: contextualizando el comentario (segmento narrativo), presentando las características del producto (segmento descriptivo) y, finalmente, procurando persuadir al lector para que adquiera o rechace el producto (segmento valorativo). A nivel de registro, los comentarios se caracterizan por el uso de un registro lingüístico marcadamente coloquial. El Cuadro 32 resume estas propiedades del, en adelante, género discursivo de los comentarios.

GÉNERO DISCURSIVO	comentario (*review*)
TIPOS DE SEGMENTOS	secuencias narrativas, descriptivas y valorativas
PROPÓSITO COMUNICATIVO GENERAL	valorar positiva o negativamente un producto
PROPÓSITO COMUNICATIVO SEGÚN SEGMENTO	narrativo: contextualizar el comentario descriptivo: presentar el producto valorativo: persuadir al lector
REGISTRO	coloquial

Cuadro 32: Caracterización del género discursivo de los comentarios sobre productos.

En el siguiente capítulo me apoyaré en esta caracterización del género discursivo de los comentarios para calcular su polaridad.

5 ANÁLISIS DE LA POLARIDAD DE LOS COMENTARIOS

Resumen

En este capítulo analizo la polaridad de los comentarios sobre productos. El objetivo de este análisis es determinar la función que cumplen los segmentos narrativo, descriptivo y valorativo en la expresión de la polaridad. Los experimentos que presento a lo largo del capítulo están destinados a evaluar tres métodos alternativos para identificar de manera automática los segmentos discursivos y a determinar si su uso facilita la detección de la polaridad de los comentarios. Un primer método clasifica los tres tipos de segmentos a partir del uso de información morfosintáctica y léxica. Una vez identificados, los segmentos se representan mediante una bolsa de palabras usando diferentes esquemas de pesado y se aplica cada segmento por separado para determinar la polaridad de los comentarios. Un segundo método extrae las «secuencias narrativas» de los comentarios y contrasta su rendimiento con la de otros fragmentos del texto para calcular la polaridad del comentario. Un tercer método aplica la «complejidad sintáctica» para clasificar las oraciones descriptivas que expresan la misma polaridad (oraciones simétricas) o la polaridad opuesta (oraciones asimétricas) a la del comentario. Posteriormente, se determina el impacto que tiene la omisión de las oraciones asimétricas en el cálculo de la polaridad del comentario. Los resultados obtenidos en esta parte de la investigación aparecen publicados en:

- John Roberto, Maria Salamó y M. Antònia Martí [2015a], «Genre-Based Stages Classification for Polarity Analysis», en *The 28th Florida Artificial Intelligence Society Conference (FLAIRS), USA*, vol. 1, pág. 1-6
- John Roberto, Maria Salamó y M. Antònia Martí [2015b], «Polarity analysis of reviews based on the omission of asymetric sentences», *Procesamiento del Lenguaje Natural*, 54, pág. 77-84
- John Roberto, Maria Salamó y M. Antònia Martí [2014], «The function of narrative chains in the polarity classification of reviews», *Procesamiento del Lenguaje Natural*, 52, pág. 69-76

5.1 INTRODUCCIÓN

En el capítulo anterior se estableció que el acto de valorar positiva o negativamente un producto, bien o servicio es un evento comunicativo complejo que, desde el punto de vista discursivo, involucra el uso de varios tipos de segmentos, cada uno de los cuales ha de cumplir una función específica en la expresión de la polaridad. En concreto, se determinó que el comentario está formado por tres tipos de segmentos: un segmento narrativo, orientado a contextualizar la opinión vertida en el comentario; un segmento descriptivo, destinado a presentar las características del producto; y un segmento valorativo, mediante el cual el autor del comentario persuade al lector para que adquiera o rechace un producto.

Partimos de la hipótesis que cada tipo de segmento tiene un peso distinto en la expresión de la polaridad del comentario. En este capítulo presento

varios experimentos orientados a identificar de forma automática cada tipo de segmento y a evaluar el rendimiento, en términos de precisión, que cada uno de ellos presenta para predecir la polaridad de los comentarios. Al contrastar la precisión individual con la del comentario completo, será posible obtener los datos necesarios para determinar qué función cumplen los segmentos en la expresión de la polaridad. A lo largo del capítulo propongo tres métodos alternativos para calcular la polaridad de los comentarios mediante el uso de los tipos de segmentos discursivos propuestos. La selección de estos tres métodos de clasificación obedece a la necesidad de tratar cada tipo de segmento según el propósito comunicativo que le caracteriza.

El primer método selecciona un conjunto de rasgos lingüísticos que utilizo como atributos de entrenamiento para realizar una clasificación supervisada de los tres tipos de segmentos. Estos rasgos describen algunas de las propiedades morfosintácticas y léxicas más características de cada tipo de segmento. Aplicando una aproximación supervisada basada en el uso de bolsa de palabras (BoW) y otra no supervisada basada en la herramienta SO-CAL, contrasto el rendimiento que cada tipo de segmento presenta al ser usado para calcular la polaridad del comentario completo. Los resultados obtenidos mediante este primer análisis me han permitido identificar la función que desempeña el segmento valorativo en la expresión de la polaridad.

El segundo método se centra en identificar las secuencias narrativas que componen el comentario. Para detectar dichas secuencias narrativas, e inspirado en los trabajos de Chambers [2011], implemento un algoritmo que extrae las oraciones que en un comentario «narran» eventos relacionados temporalmente, es decir, las oraciones que conformarán las secuencias narrativas. Una vez recuperadas estas secuencias, realizo varios experimentos encaminados a determinar el impacto que su omisión tiene a nivel del cálculo de la polaridad de los comentarios. Los resultados obtenidos mediante este segundo análisis me han permitido identificar la función que desempeña el segmento narrativo en la expresión de la polaridad.

Finalmente, el tercer método recupera las oraciones del comentario que describen las características positivas y negativas de un producto (ej. «un airbag de conductor con una forma optimizada para proporcionar una mayor eficacia»), es decir, el segmento descriptivo. Aplicando aprendizaje supervisado, clasifico estas oraciones como simétricas o asimétricas, según expresen o no la misma polaridad que la del comentario. Para el entrenamiento de estos clasificadores utilizo como atributos diferentes índices de la complejidad sintáctica como son los índices de Yngve [Yngve, 1960], Frazier [Frazier y Clifton, 1998] y Pakhomov [Pakhomov y col., 2011]. Posteriormente, realizo una serie de experimentos orientados a determinar el impacto que la omisión de las oraciones descriptivas asimétricas tiene sobre la polaridad de los comentarios. Los resultados obtenidos mediante este tercer análisis me han permitido identificar la función que desempeña el segmento descriptivo en la expresión de la polaridad.

En la Sección 5.2 presento los experimentos destinados a identificar de manera automática los tres tipos de segmentos discursivos que componen los comentarios; en la Sección 5.3, los experimentos que determinan el rendimiento de cada tipo de segmento al ser usado para calcular de la polaridad de los comentarios; en la Sección 5.4, los experimentos orientados a identificar y aplicar las secuencias narrativas al cálculo de la polaridad; y en la Sección 5.5, los experimentos que clasifican y aplican las oraciones simétricas y asimétricas al cálculo de la polaridad. Adicionalmente, en la Sección 5.6 establezco la función que desempeña cada segmento en la expresión de

la polaridad a la luz de los experimentos descritos a lo largo del capítulo. Finalmente, en la Sección 5.7 presento las conclusiones que se derivan del análisis de la polaridad de los comentarios.

5.2 IDENTIFICACIÓN AUTOMÁTICA DE LOS SEGMENTOS DISCURSIVOS

En esta sección presento los experimentos llevados a cabo para identificar de forma automática los tres tipos de segmentos que forman parte del género discursivo de los comentarios mediante el uso de información morfosintáctica y léxica. En los experimentos aplico técnicas de aprendizaje supervisado. La tarea consiste en predecir si un segmento x pertenece a alguna de las siguientes clases: narrativo, descriptivo o valorativo. Para la fase de entrenamiento de los clasificadores, se han anotado manualmente los tres tipos de segmentos en un conjunto de comentarios en español y en inglés. En total he anotado 150 comentarios sobre hoteles provenientes del corpus HOpinion y 90 comentarios sobre coches del corpus MDTOD [Cruz Mata, 2012]. Como se aprecia en el Cuadro 33, la longitud de los textos de HOpinion oscila entre las 28 y las 958 palabras, con una media de 179 palabras por comentario y un total de 35.053 palabras. La longitud de los textos del MDTOD está entre las 20 y las 2046 palabras, con una media de 326 palabras por comentario y un total de 38.365 palabras. El formato de anotación de los textos es XML. En el apartado siguiente describo el esquema general de anotación.

Corpus	Idioma	Dominio	Muestras	Min. Pals.	Max. Pals.	Media Pals.	Total Pals.
HOpinion	Español	hoteles	150	28	958	179	35.053
MDTOD	Inglés	coches	90	20	2046	326	38.365

Cuadro 33: Resumen de los datos empleados para validar la presencia de los segmentos narrativos, descriptivos y valorativos en los comentarios sobre productos.

5.2.1 Esquema general de anotación

El esquema general de anotación describe las características morfosintácticas y léxicas que se usarán para clasificar de forma automática el contenido de los comentarios en tres clases de segmentos: valoración, narración y descripción. Para seleccionar estas características me baso en los trabajos existentes sobre tipologías textuales y en la observación directa de los textos. En este último caso, el análisis y la caracterización de los comentarios que desarrollé en el Capítulo 4 me han permitido establecer ciertas diferencias en cuanto a la composición de los tres tipos de segmentos. Antes de presentar el esquema general de anotación que he aplicado a los comentarios, comentaré brevemente algunas de las propiedades más representativas de cada segmento.

En primer lugar, los segmentos presentan diferencias en cuanto al número de palabras o unidades léxicas que los componen: los segmentos narrativo y descriptivo suelen ser más extensos que el valorativo. Además, el uso del tiempo futuro, el modo condicional y el subjuntivo representan mejor al

segmento valorativo (ej. «no se lo recomendaría a nadie»). En cuanto a la persona gramatical, los trabajos existentes sobre tipologías textuales suelen relacionar los textos narrativos con el uso de la primera persona (ej. «pasamos dos días agradables»), las descripciones con la tercera persona (ej. «el hotel está en una ubicación muy apropiada») y las valoraciones con la primera o segunda persona (ej. «no me plantearía alojarme en ningún otro sitio», «*you should certainly consider other options*»).

En segundo lugar, los estudios sobre los modelos textuales han determinado que ciertas clases de verbos son más frecuentes en las narraciones que en las descripciones. De acuerdo con Schilder [1997] las clases eventivas, en cuanto cualidades temporales propias de la situación designada por un verbo, suelen estar vinculadas a la expresión de narraciones. En la misma línea, M. Carroll y Lambert [2003] demostraron que tanto el aspecto gramatical como el léxico (*Aktionsart*) están relacionados con la forma en que se estructuran los textos narrativos. Por último, cada tipo de segmento se caracteriza por el uso de determinadas unidades léxicas asociadas al propósito comunicativo que le es propio. Por ejemplo, los marcadores discursivos que indican resumen o conclusión tienen razón de ser en los segmentos valorativos (ej. «en fin», «de todos modos...», «por último», «*anyway*», «*finally*»). Las expresiones temporales son, por definición, características de los segmentos narrativos (ej. «hace * años», «ayer», «*the day before*», «*in the morning*»). En los segmentos descriptivos son recurrentes los adverbios y adjetivos. Además, al tratarse de comentarios sobre productos, se encuentran términos de dominio formando parte del segmento descriptivo. Por ejemplo, en el dominio de los coches son frecuentes términos como: «motor», «válvula», «diésel», «*spare wheel*», «*wheelbase*».

El esquema general de anotación utilizado para enriquecer los comentarios sobre productos en español e inglés con esta y otra información lingüística adicional, es el siguiente:

SEGMENTOS DISCURSIVOS: Los comentarios fueron segmentados y anotados manualmente con los atributos de clase, es decir, con los segmentos narrativos, descriptivos y valorativos, atendiendo a los criterios establecidos en la Sección 4.2.2. A continuación presento, a modo de ejemplo, el caso de un segmento valorativo tomado del corpus HOpinion y que coincide con la frase «No tardaré en volver». Como se puede ver en el ejemplo, el tipo de segmento o atributo de clase (variable a predecir) está anotado mediante la etiqueta *stage type*.

```
<stage type="valorative">
    <r lem="no" name="r" pos="rn" postype="negative" wd="No"/>
    <v lem="tardar" lexaspect="A23-C31" mood="indicative" name="v" num="s"
        person="1" pos="vmif1so" postype="main" tense="future" wd="tardaré"/>
    <s complex="no" lem="en" name="s" pos="spsoo" postype="preposition" wd="en
        "/>
    <v lem="volver" lexaspect="B11-A11-A31-D11" mood="infinitive" name="v" pos
        ="vmnoooo" postype="main" wd="volver"/>
    </stage>
```

PART-OF-SPEECH (POS): Puesto que los comentarios sobre productos presentan un número importante de rasgos comunicativos y lingüísticos que no son característicos del español o del inglés estándar, los comentarios fueron lematizados. La lematización consiste en asociar a cada unidad léxica su forma canónica o lema correspondiente. Esta información contribuye a la normalización de los textos puesto que los lemas

son formas estandarizadas. Adicionalmente, los comentarios fueron etiquetados con información morfosintáctica (*Part-Of-Speech tagging*) que fue verificada manualmente por un lingüista. Este procedimiento se utiliza para asignar a cada unidad léxica una única categoría morfosintáctica. Como se puede observar en el siguiente ejemplo, este análisis permite segmentar los comentarios en sus unidades léxicas para obtener, entre otras informaciones: el lema, el modo, el tiempo, la persona y la categoría gramaticales.

```
<v lem="tardar" name="v" num="s" mood="indicative" tense="future" person="1"
    pos="vmif1so" postype="main" wd="tardaré"/>
```

MARCADORES DISCURSIVOS: Se etiquetaron los marcadores discursivos que el autor del comentario utiliza para resumir, concluir o reformular su opinión. Algunos ejemplos de estos marcadores o conectores discursivos son: «en conclusión», «en resumen», «*finally*», «*anyway*». La lista de conectores fue compilada a partir de los adverbios y locuciones adverbiales detectados en los comentarios como resultado del proceso automático de etiquetado morfosintáctico, tal y como queda representado en el ejemplo que viene a continuación.

```
<r lem="de_todos_modos" name="r" pos="rg" wd="De_todos_modos"/>
```

TÉRMINOS ESPECÍFICOS DEL DOMINIO: Se anotaron los términos específicos del dominio presentes en los comentarios (ej. «frigobar», «*staff*», «*check-in*»). En el caso de los textos de HOpinion, esta tarea se realizó construyendo un léxico propio a partir del uso de un conjunto reducido de palabras semilla (*seed words*, sw) extraídas de la ontología Hontology (ver Sección 3.3.2.5) y de los más de 18.000 comentarios sobre productos que integran el corpus HOpinion. Una vez obtenidas las palabras semilla, he seleccionado los nombres y verbos que aparecen cerca de dichas palabras dentro de una ventana de tamaño 3. Este listado de palabras conforman los candidatos a términos específicos del dominio (*domain-specific term candidates*, tc): ej. «El número máximo de$_3$ **huéspedes**$^{tc}_{2}$ por$_1$ **habitación**$^{sw}_{0}$ es$_1$ de$_2$ cuatro$_3$» . El listado de candidatos a términos específicos del dominio fue revisado manualmente antes de ser incluido como un nuevo recurso. En total, el recurso contiene 784 términos, incluyendo algunas variantes de una misma palabra como, por ejemplo, «refrigerador», «refrigeradora», «frigorífico», «frigo», «frigobar», «mini-bar», «minibar». Un procedimiento más simple se usó en el dominio de los coches en inglés puesto que en este caso el corpus MDTOD ya viene acompañado con una taxonomía de términos de dominio (ver Apartado 3.2.2). A continuación, un ejemplo de término de dominio con su información morfosintáctica asociada.

```
<n gen="f" lem="habitación" domain="yes" name="n" num="p" pos="ncfp000" postype
    ="common" wd="habitaciones"/>
```

TÉRMINOS ESPECÍFICOS DE CADA TIPO DE SEGMENTO: Se anotaron los términos específicos de cada tipo de segmento o palabras clave (*keywords*) constituidos por unigramas y trigramas. En el análisis no empleo bigramas o n-gramas de más de tres palabras puesto que: **1.** Bekkerman y Allan [2003] demostraron que los bigramas presentan serias limitaciones aplicados a tareas de clasificación de textos, y **2.** Cui y col. [2006] constataron que los n-gramas complejos (*high order n-grams* $= n > 3$)

son poco efectivos en tareas de minería de emociones y de sentimientos. Los trigramas, por el contrario, son lo suficientemente extensos como para codificar información sintáctica relevante y, al mismo tiempo, lo suficientemente pequeños como para poder operar con ellos a nivel computacional. En el siguiente ejemplo aparece el trigrama *dudosamente_el_recomendar* («dudosamente lo recomendaría»), característico del segmento valorativo.

```
<stage type="valorativo">
...
<r lem="dudosamente" name="r" pos="rg" wd="dudosamente"/>
<d gen="n" lem="el" name="d" num="s" pos="daonso" postype="article" wd="lo"/>
<v lem="recomendar" mood="indicative" name="v" num="s" person="1" pos="vmic1so
    " postype="main" tense="conditional" wd="recomendaría"/>
...
```

ASPECTO LÉXICO DE LOS VERBOS (*aktionsart*): Siguiendo la clasificación de los verbos propuesta por Aparicio y col. [2008], los verbos de los comentarios fueron anotados según su aspecto léxico, modo de acción o *Aktionsart* (estados, realizaciones, actividades y logros). La clasificación y anotación de los verbos se realizó de forma automática a partir del léxico verbal AnCoraVerbES (ver Apartado 3.3.2.2). El ejemplo que viene a continuación presenta un verbo («estar») anotado con su clase eventiva correspondiente (C=estado)[1].

```
<v lem="estar" lexaspect="C11–C21" mood="indicative" name="v" num="s" person
    ="1" pos="vmis1so" postype="main" tense="past" wd="Estuve"/>
```

A partir de la información descrita en los párrafos anteriores, he seleccionado un conjunto de rasgos o características que considero relevantes para diferenciar los tres tipos de segmentos que integran los comentarios sobre productos: narración, descripción y valoración. El Cuadro 34 contiene la lista completa de estas características, 14 en total, identificadas mediante un ID que servirá para referirme a cada una de ellas en lo que queda de sección.

5.2.2 Segmentación de comentarios en español

En esta sección presento el experimento que he llevado a cabo para clasificar los segmentos narrativo, descriptivo y valorativo en los 150 comentarios sobre hoteles en español extraídos del corpus HOpinion. La clasificación se realiza mediante aprendizaje supervisado empleando como atributos las 14 características lingüísticas seleccionadas en el apartado anterior (ver Cuadro 34).

Para poder usar estas características, los comentarios del corpus se han de transformar en una representación adecuada que permita aplicar los diferentes algoritmos de aprendizaje. A continuación formalizo la representación de los comentarios del corpus HOpinion. De una parte, tenemos un conjunto S con 150 comentarios compuesto por tres tipos de segmentos (s_n = narrativo, s_d = descriptivo y s_v= valorativo) de manera que $S = \{s_n, s_d, s_v\}$. De otra parte, contamos con un conjunto de rasgos $F = \{f_1 ... f_{14}\}$, que caracterizan los elementos (segmentos) de S, y una matriz M en la que los elementos

1 Según la nomenclatura utilizada para la generación del AnCoraVerbES, la letra (A, B, C, D) identifica la clase eventiva y los números (11, 21, ...) especifican las alternancias diatéticas en que puede aparecer cada verbo, es decir, el número de argumentos requeridos por el verbo (valencia del verbo).

Característica	ID
Número de palabras	f_1
Tiempo y modo	f_2
Primera persona	f_3
Segunda persona	f_4
Tercera persona	f_5
Términos específicos del dominio	f_6
Términos específicos del tipo de segmento (unigramas)	f_7
Términos específicos del tipo de segmento (trigramas)	f_8
Aktionsart (realización)	f_9
Aktionsart (logro)	f_{10}
Aktionsart (estado)	f_{11}
Aktionsart (actividad)	f_{12}
Verbos (frecuencia)	f_{13}
Marcadores discursivos	f_{14}

Cuadro 34: Características morfosintácticas y léxicas empleadas para diferenciar los segmentos narrativos, descriptivos y valorativos en los corpus HOpinion y MDTOD.

de F conforman las columnas y los de S las filas: $M = \{S \text{ x } F\}$. Puesto que el número máximo de segmentos en S es 450 (150 comentarios x 3 posibles tipos de segmentos), entonces $S = \{s_1 ... s_{\leqslant 450}\}$. Hago notar que, en esta representación, los segmentos con una misma etiqueta se han contabilizado conjuntamente. Por ejemplo, en la secuencia `D – N – D`, se contabilizan 2 tipos de segmentos, no tres. Adicionalmente, he realizado una transformación lineal de los datos originales para ubicar todos los valores de M en el rango [0..1]. La normalización de los valores de la matriz M se calculan mediante la fórmula (1), donde, m_{ij} es el valor actual del ejemplo i para la característica j, $min(f_j)$ es el valor mínimo para f_j en F ($f_j \in F$) y $max(f_j)$ el valor máximo.

$$Norm(m_{ij}) = \frac{m_{ij} - min(f_j)}{max(f_j) - min(f_j)} \tag{1}$$

Para la clasificación, mediante el entorno Weka [Witten y col., 1999], he agrupado los segmentos (sg) en una clase ternaria y dos clases binarias (de elementos disjuntos $\veebar$): narración versus descripción versus valoración (N $\veebar$ D $\veebar$ V), narración versus descripción (N $\veebar$ D), narración versus valoración (N $\veebar$ V) y descripción versus valoración (D $\veebar$ V). En total, he aplicado 19 algoritmos de aprendizaje (ag), dos métodos de selección de atributos (ms) y un método de búsqueda (mb) (ver Cuadro 35). Los datos fueron divididos en un conjunto de entrenamiento y uno de prueba utilizando validación cruzada (10-fold cross-validation). El número de clasificadores entrenados es, por tanto, de 1.520: 4_{sg} x 19_{ag} x 2_{ms} x 1_{mb} x 10_{folds}.

El resultado de la clasificación de los segmentos aparece en el Cuadro 36. La primera columna muestra las cuatro configuraciones en las que se pueden combinar los tipos de segmentos. La segunda columna contiene el promedio de precisión o exactitud obtenido en cada configuración. La columna tres ofrece las características más relevantes recogidas por los métodos de selección de atributos en cada una de las cuatro configuraciones. En esta colum-

na, el asterisco significa que dos características son igualmente relevantes. La cuarta columna presenta el identificador de cada uno de los rasgos de acuerdo con lo establecido en el Cuadro 34.

Algoritmos de aprendizaje (ag) **1. Bayes:** BayesNet y DMNBtext **2. Perezosos:** IBk, KStar y LWL **3. Reglas:** ConjunctiveRule, DTNB, DecisionTable, JRip, OneR y PART **4. Árboles:** ADTree, BFTree, J48, J48graft, LMT, NBTree, REPTree y RandomForest)
Métodos de selección (ms) Information Gain (IG) y χ^2 (CHI-SQUARE)
Método de búsqueda (mb) Ranker

Cuadro 35: Lista de algoritmos, métodos de selección y búsqueda utilizados para clasificar los segmentos narrativo, descriptivo y valorativo.

Segmentos	Precisión	Características	#f
N $\veebar$ D $\veebar$ V	81.4 %	Términos específicos del tipo de segmento (unigramas)	f_7
		Aktionsart (actividad)*	f_{12}
		Términos específicos del dominio*	f_6
N $\veebar$ D	97.4 %	Términos específicos del tipo de segmento (unigramas)	f_7
		Aktionsart (actividad)*	f_{12}
		Términos específicos del dominio*	f_6
N $\veebar$ V	83.7 %	Términos específicos del tipo de segmento (unigramas)*	f_7
		Frecuencia de los verbos*	f_{13}
		Primera persona	f_3
D $\veebar$ V	93.1 %	Términos específicos del dominio	f_6
		Primera persona	f_3
		Tercera persona	f_5
Promedio	**88.9 %**		

Cuadro 36: Resultado de la clasificación de los segmentos en el corpus HOpinion.

Los resultados indican que el mejor rendimiento se obtiene cuando las clases a aprender son la narrativa frente a la descriptiva, es decir, la configuración N $\veebar$ D. La exactitud promedio en este caso es del 97.4 %. Esto se debe a que el segmento narrativo y el descriptivo tienen funciones comunicativas bien diferenciadas: contextualizar el comentario y describir el producto. También se obtuvo un excelente rendimiento al contrastar los segmentos descriptivo y valorativo. La exactitud promedio conseguida con los 380 clasificadores entrenados en esta configuración (D $\veebar$ V) es de 93.1 %. Este porcentaje no es tan elevado como el de la configuración anterior puesto que en algunos casos el segmento valorativo se usa para resumir el comentario y, por lo tanto, comparte algunas de las características con el segmento descriptivo como, por ejemplo, el vocabulario. El hecho de que los «términos

específicos del tipo de segmento» (f_7) no sea, en este caso, la característica predominante, refuerza esta conclusión.

Las otras dos configuraciones, N $\veebar$ V y N $\veebar$ D $\veebar$ V, no obtuvieron precisiones tan altas como las de las clases binarias compuestas por un segmento descriptivo. No obstante, dicho rendimiento supera, en ambos casos, el 80 %. La exactitud promedio alcanzada al diferenciar los segmentos narrativo y valorativo (N $\veebar$ V) es del 83.7 %. Estos resultados indican que el segmento descriptivo es el de más fácil identificación. Ya en el Capítulo 4 sostenía que las descripciones son un segmento preceptivo en los comentarios sobre productos. La exactitud promedio obtenida en el caso de la configuración N $\veebar$ D $\veebar$ V es de 81.4 %, la más «baja» de las cuatro. Esto se debe, evidentemente, a que se trata de una clase tripartita, es decir, un problema de aprendizaje multiclase.

El resultado en cuanto a la relevancia de las características es el siguiente. En todas las configuraciones, excepto D $\veebar$ V, los «términos específicos del tipo de segmento» (f_7) es la característica más relevante para diferenciar los tipos de segmentos. La configuración D $\veebar$ V no utiliza este rasgo porque, en un número importante de ocasiones, las valoraciones se combinan con las descripciones para ofrecer una justificación: «no pienso volver porque *el personal es desagradable*»). Adicionalmente, los «términos específicos del dominio» han resultado de gran utilidad para caracterizar al segmento descriptivo, el rasgo f_6 aparece en todas las configuraciones que involucran una descripción: N $\veebar$ **D** $\veebar$ V, N $\veebar$ **D** y **D** $\veebar$ V. Tal y como comenté en el Apartado 4.2.2.2, a diferencia de los segmentos narrativo y valorativo, el segmento descriptivo se caracteriza por contener información sobre el dominio.

Los verbos son relevantes para identificar los segmentos narrativos. Este tipo de segmento aporta información sobre el usuario a través de pequeños relatos: «Esta **ha sido** mi tercera vez **visitando** Barcelona y en todas **he elegido** el mismo hotel porque ...». En concreto, los verbos de actividad (f_{12}) resultan de utilidad para diferenciar los segmentos narrativos de los descriptivos. Este rasgo tiene la misma relevancia que los «términos específicos del dominio» (f_7) en las configuraciones **N** $\veebar$ **D** $\veebar$ V y **N** $\veebar$ **D**. De manera similar, la frecuencia de los verbos (f_{13}) permite diferenciar el segmento narrativo del valorativo. En general, los dos rasgos relacionados con el uso de los verbos (f_{12} y f_{13}) aparecen en las configuraciones que involucran un segmento narrativo: **N** $\veebar$ D $\veebar$ V, **N** $\veebar$ D y **N** $\veebar$ V.

Por último, el uso de la «primera persona» (f_3) caracteriza al segmento valorativo. Dos de las tres configuraciones compuestas por un segmento valorativo dan relevancia a este rasgo: N $\veebar$ **V** y D $\veebar$ **V**. Como comenté en el Apartado 4.2.2.3, los segmentos valorativos describen la actitud personal del usuario respecto del producto valorado. En esta clase de segmento, el usuario realiza una reflexión personal en favor o en contra del producto, de ahí la importancia que tiene el uso de la «primera persona».

En general, los resultados obtenidos demuestran que es posible clasificar los segmentos narrativo, descriptivo y valorativo de los comentarios sobre productos en español (corpus HOpinion) con un rendimiento superior al 80 %.

5.2.3 Segmentación de comentarios en inglés

En esta sección presento el experimento que he llevado a cabo para clasificar los segmentos narrativo, descriptivo y valorativo en los 90 comentarios sobre coches en inglés extraídos del corpus MDTOD. La clasificación

se realiza mediante aprendizaje supervisado empleando como atributos las 14 características lingüísticas descritas en el esquema general de anotación (Apartado 5.2.1, Cuadro 34).

La configuración de este experimento es la misma que se ha utilizado para clasificar los segmentos en español (ver apartado 5.2.2): misma representación de los comentarios, misma herramienta (Weka), mismos algoritmos de aprendizaje, métodos de selección y búsqueda. De igual manera, los datos fueron divididos en un conjunto de entrenamiento y uno de prueba utilizando validación cruzada (10-fold cross-validation). Los resultados de la segmentación automática aparecen en el Cuadro 37.

Segmentos	Exactitud	Características	#f
N $\veebar$ D $\veebar$ V	78.5 %	Términos específicos del tipo de segmento (unigramas)	f_7
		Primera persona	f_3
		Aktionsart (logro)	f_{10}
N $\veebar$ D	82.1 %	Términos específicos del tipo de segmento (unigramas)	f_7
		Primera persona	f_3
		Tercera persona	f_5
N $\veebar$ V	87.9 %	Términos específicos del tipo de segmento (unigramas)	f_7
		Aktionsart (realización)	f_9
		Aktionsart (logro)	f_{10}
D $\veebar$ V	91.6 %	Términos específicos del tipo de segmento (unigramas)	f_7
		Tercera persona	f_5
		Aktionsart (realización)	f_9
Promedio	**84.7 %**		

Cuadro 37: Resultado de la clasificación de los segmentos en el corpus MDTOD.

Tal y como se puede observar, el funcionamiento de los clasificadores para el dominio de coches en inglés es muy similar al conseguido para el dominio de los hoteles en español (ver Cuadro 36). Esto significa que los segmentos tienden a comportarse y mantener las mismas propiedades en ambos dominios e idiomas. El promedio de las cuatro configuraciones analizadas para el corpus HOpinion es de 88.9 % y para el corpus MDTOD es de 84.7 %. La mayor diferencia entre ambos corpus involucra la configuración N $\veebar$ D: HOpinion=94.4 y MDTOD=82.1. Las otras tres configuraciones, que utilizan un segmento de tipo valorativo, presentan niveles de precisión muy similares, especialmente D $\veebar$ V: HOpinion=93.1 y MDTOD=91.6. La configuración N $\veebar$ V mejora su rendimiento con el cambio de dominio e idioma: HOpinion=83.7 y MDTOD=87.9.

Los métodos de selección de rasgos identificaron los «términos específicos del tipo de segmento» (f_7) como la característica más relevante para clasificar los tipos de segmentos. Este rasgo también ha sido fundamental para clasificar los comentarios de HOpinion. Contrariamente a lo que sucede con los comentarios de hoteles en español, los «términos específicos del dominio» (f_6) no sirven para discriminar los segmentos en el corpus de coches en inglés. Finalmente, como se desprende del uso de las características, los

(tipos de) verbos (f_9 y f_{10}) cumplen un función esencial para diferenciar los segmentos narrativos de los valorativos en inglés.

Los resultados obtenidos demuestran que es posible clasificar los segmentos narrativo, descriptivo y valorativo de los comentarios sobre productos en inglés (corpus MDTOD) con una precisión promedio del 84.7 %, esto es, una precisión similar a la obtenida para los comentarios en español. Por lo tanto, la siguiente fase del análisis consiste en usar estos segmentos para clasificar los comentarios según su polaridad. El rendimiento que obtenga cada tipo de segmento en esta tarea de clasificación será un indicador de la función que cumplen en la expresión de la polaridad.

5.3 APLICACIÓN DE LOS SEGMENTOS AL CÁLCULO DE LA POLARIDAD

En este apartado presento los experimentos que he efectuado con el fin de determinar el rendimiento que presentan los segmentos narrativo, descriptivo y valorativo para calcular la polaridad de los comentarios sobre productos. En el Apartados 5.3.1 describo los experimentos relacionados con el cálculo de la polaridad de comentarios en español y en el Apartado 5.3.2 con el cálculo de la polaridad en inglés.

5.3.1 Cálculo de la polaridad de comentarios en español

Para el cálculo de la polaridad en español he empleado 120 de los 150 comentarios obtenidos de HOpinion: $C = \{c_1 ... c_{120}\}$. Los 30 textos omitidos corresponden a los comentarios con una valoración de 3 puntos/estrellas, es decir, los comentarios «neutros». A diferencia de las investigaciones orientadas al tratamiento de la intensidad de las opiniones (*strength of opinions*), la supresión de los comentarios neutros es un procedimiento habitual en el análisis de la polaridad (*polarity of opinions*) mediante técnicas de aprendizaje automático. Los textos seleccionados han sido modelados mediante una matriz R, en la que las filas contienen los comentarios y las columnas las palabras (*words*, *W*), de manera que $R = \{C \times W\}$. Cada elemento de C, es decir $c_k \in C$, es representado por un conjunto finito de palabras o bolsa de palabras (*Bag of Words*, BoW). Al mismo tiempo, a cada comentario C, le corresponden tres representaciones posibles de BoW, de acuerdo con los tres tipos de segmentos S definidos: segmento narrativo $BoW_n = \{w_{n1} ... w_{nt}\}$, segmento descriptivo $BoW_d = \{w_{d1} ... w_{dt}\}$ y segmento valorativo $BoW_v = \{w_{v1} ... w_{vt}\}$. Adicionalmente, con el fin de contar con una representación que funcione como marco de referencia, cada comentario C se ha representado mediante una cuarta bolsa de palabras (BoW_a) que contiene todos los datos disponibles, esto es, las tres representaciones de C. Por lo tanto, $BoW_a = \{w_{a1} ... w_{at'}\}$ donde $t' = |BoW_n| + |BoW_d| + |BoW_v|$.

Por último, he aplicado tres esquemas de pesado para determinar la importancia de las palabras w_i en cada una de las representaciones del tipo BoW. Los tres esquemas utilizados son el *booleano* o *binario* (`bin`), la *frecuencia absoluta* (`tf`) y la *frecuencia relativa* (`tf-idf`). En primer lugar, el esquema *binario* es simple y sólo considera la aparición o no de términos en el documento. Este esquema de pesado asigna un valor de «1» al término si éste aparece en el comentario, y «0» en caso contrario: $w_i \in \{0, 1\}$. En segundo lugar, según la *frecuencia absoluta* del término o `tf` (*Term Frecuency*), el peso

del término w_i en el comentario c_j se corresponde con la cantidad de veces que aparece w_i en c_j y se denota por $tf(w_i, c_j)$. En tercer lugar, la *frecuencia relativa* es el cociente entre la frecuencia absoluta de un término y el número total de datos. En otras palabras, en este enfoque se determina el peso de un término w_i en el comentario c_j en proporción directa al número de veces que el término aparece en el comentario, e inversamente proporcional al número de comentarios en los que aparece el término en el conjunto total de entrenamiento C. El peso de `tf-idf` viene dado por la fórmula 2, donde f_{ik} es la frecuencia absoluta de w_i en el comentario c_j, es decir, $f_{ik} = tf(w_i, c_j)$; C es el número de comentarios de la colección y c_k el número de comentarios en los que w_i aparece, es decir, $c_k \in C : w_i \in c_k$. Este método da mayor relevancia a los términos que ocurren frecuentemente en un documento pero que son poco frecuentes en la colección.

$$p(w_{ik}) = f_{ik} x \log\left(\frac{C}{c_k}\right) \tag{2}$$

En este experimento he evaluado 37 algoritmos de aprendizaje agrupados en seis grandes categorías. La lista detallada de algoritmos aparece en el Cuadro 38. Nuevamente, cada algoritmo fue evaluado empleando validación cruzada de diez iteraciones (10-fold cross-validation). En total, he construido 4.440 modelos para este análisis: 10 iteraciones x 37 algoritmos x 3 esquemas de pesado x 4 representaciones BoW.

Algoritmos de aprendizaje

1. **Bayes:** BayesNet, BayesianLogisticRegression, ComplementNaiveBayes, DMNBtext, NaiveBayes, NaiveBayesMultinomial, NaiveBayesMultinomialUpdateable y NaiveBayesUpdateable.
2. **Perezosos:** IB1, IBk, KStar y LWL.
3. **Misceláneos:** HyperPipes y VFI.
4. **Reglas:** ConjunctiveRule, DTNB, DecisionTable, JRip, NNge, OneR, PART, Ridor y ZeroR.
5. **Árboles:** ADTree, BFTree, DecisionStump, FT, J48, J48graft, LADTree, NBTree, REPTree, RandomForest, RandomTree, SimpleCart y lmt.LogisticBase.
6. **Funciones:** SMO.

Cuadro 38: Lista de algoritmos utilizados para clasificar la polaridad de los comentarios en el corpus HOpinion.

En el Cuadro 39 aparecen los resultados conseguidos al clasificar los comentarios según su polaridad mediante las cuatro representaciones analizadas: $S_n = BoW_n$ (segmentos narrativos), $S_d = BoW_d$ (segmentos descriptivos), $S_v = BoW_v$ (segmentos valorativos) y $S_a = BoW_a$ (todos los segmentos juntos). Los valores expuestos en la parte superior del Cuadro 39 corresponden a la **precisión** obtenida al aplicar los diferentes esquemas de pesado a cada una de las representaciones vectoriales en las que se han transformado los comentarios. Cada uno de estos valores muestra la precisión promedio conseguida con los 370 clasificadores que resultan de cruzar la representación de los segmentos con los pesos de los términos. Según los resultados, sólo S_a y S_v alcanzan precisiones superiores al 70 %. El mejor resultado lo obtuvo la representación S_a usando `tf` y `tf-idf` como esquemas de pesado: $S_a/tf = 74.1\,\%$ y $S_a/tf - idf = 73\,\%$. En el caso del segmento

valorativo, la mejor combinación es para $S_v/\mathtt{bin} = 70.9\,\%$ y $S_v/\mathtt{tf} = 70.6\,\%$. Dado que los segmentos valorativos son más breves y contienen menos palabras que los otros dos segmentos, es normal que el esquema de pesado que represente mejor su rendimiento sea el binario y el que peor lo represente sea la frecuencia relativa de los términos, $S_v/\mathtt{tf}-\mathtt{idf} = 68.5\,\%$. De acuerdo con los **promedios de precisión**[2], las diferentes representaciones se pueden ordenar según su relevancia para predecir la polaridad de los comentarios: $S_a/\mathtt{tf} > S_v/\mathtt{bin} > S_d/\mathtt{tf}-\mathtt{idf} > S_n/\mathtt{tf}$.

	Representación de los comentarios			
Esquemas de pesado	S_n	S_d	S_v	S_a
`bin`	64.39 %	63.53 %	**70.96 %**	67.81 %
`tf`	66.08 %	65.09 %	**70.63 %**	**74.18 %**
tf-idf	63.44 %	68.91 %	68.51 %	**73.08 %**
promedios de precisión	64.63 %	65.84 %	70 %	71.69 %
Cantidad de texto	23.98 %	65.21 %	10.81 %	100 %
IRA	2.7	1.0	6.5	0.7
IRR	19.42	34.37	3.48	0

Cuadro 39: Precisión promedio obtenida al clasificar la polaridad de los comentarios sobre hoteles en español mediante cuatro representaciones diferentes: segmento narrativo, descriptivo, valorativo y comentario completo.

Además de la precisión –es decir, el número de instancias correctamente clasificadas–, la **cantidad de texto** (número de palabras) usada en cada representación es una variable importante para determinar el tipo de segmento que ofrece un mejor rendimiento para calcular la polaridad de los comentarios. Por ello, en la parte intermedia del Cuadro 39 presento la cantidad de texto que se emplea en cada representación. Así, mientras que el comentario completo, es decir S_a, emplea el 100 % del texto, el segmento valorativo o S_v tan sólo usa un 10 % del texto. Por lo tanto, en un comentario de 300 palabras, la valoración se construye con sólo 30 de estas palabras. Los promedios de la columna S_v se han obtenido con ese 10 % (10.81 %, para ser exacto) del comentario y son muy similares a los conseguidos con S_a. Esto indica que los segmentos valorativos concentran gran parte de la polaridad del comentario.

Propongo denominar ÍNDICE DE RENDIMIENTO a la relación que existe entre la precisión y la cantidad de texto usados al determinar la polaridad de un comentario. Concretamente, el índice de rendimiento indicaría **la cantidad de texto necesaria para obtener un determinado nivel de precisión** en el cálculo de la polaridad a partir de un segmento S_x. El índice de rendimiento puede calcularse de dos formas diferentes, en función de si se da mayor énfasis a la cantidad de texto o a la precisión. En primer lugar, si dividimos la precisión –por ejemplo, la más elevada de cada representación– por la cantidad de texto obtenemos el ÍNDICE DE RENDIMIENTO ABSOLUTO (`IRA`). El `IRA` indica cuál de los segmentos es el que expresa mejor la polaridad con la menor cantidad de texto (ver la fórmula 3). Como se puede observar en la parte baja del Cuadro 39, el segmento valorativo es la representación que induce la mayor reducción en la cantidad de texto empleado para predecir la polaridad del comentario: $\mathtt{IRA}(S_v) = 6.5$. En segundo lugar, si utilizamos la precisión (`prec`) más alta como referencia, es decir $S_a = 74.1\,\%$, podemos

2 Promedio de los promedios.

obtener el ÍNDICE DE RENDIMIENTO RELATIVO (IRR) de cada representación S_x. IRR da mayor énfasis a la precisión, por lo que el mejor resultado es aquel que se aproxima más al IRR de referencia $\texttt{IRR}(S_a) = 0$. La fórmula para obtener el IRR se muestra en (4). Según el IRR, el $\texttt{IRR}(S_v) = 3.4$ presenta el desempeño más próximo al de referencia $\texttt{IRR}(S_a) = 0$, confirmando la importancia del segmento valorativo en el cálculo de la polaridad de los comentarios sobre productos.

$$\texttt{IRA}(S_x) = \frac{prec(S_x)}{cantidad\_texto(S_x)} \tag{3}$$

$$\texttt{IRR}(S_x) = \frac{(prec(S_a) - prec(S_x)) * cantidad\_texto(S_x)}{10} \tag{4}$$

Por último, destacaría la relación que se establece entre los esquemas de pesado de los términos, la cantidad de texto y las tres representaciones de los comentarios: el segmento descriptivo, que tiene la mayor longitud, presenta el mejor rendimiento bajo `tf-idf`; el segmento narrativo, con una extensión media, mejora su rendimiento usando `tf`; y el segmento valorativo, el de menor extensión, presenta un rendimiento superior con un esquema *binario*. El esquema *binario* es el más directo y simple de los tres esquemas de pesado lo que hace aún más atractivo el uso del segmento valorativo para calcular la polaridad.

Estos hallazgos permiten deducir que es posible obtener un nivel de rendimiento adecuado en la predicción de la polaridad utilizando solo el segmento valorativo de los comentarios en combinación con una representación binaria de los datos y bajo una aproximación simple como es la bolsa de palabras (BoW). En los experimentos he obtenido una precisión cercana al 71 % con sólo un 10.8 % de los datos. Este nivel de precisión está muy próximo al conseguido con la representación de referencia ($S_a = 74\,\%$) que, por el contrario, se basa en una matriz con una dimensionalidad y una dispersión muy altas. Este contraste queda reflejado claramente en la Figura 16, en donde se esquematiza el índice de rendimiento (`IRA` e `IRR`), es decir, la relación entre la cantidad de texto y los niveles de precisión alcanzados en cada representación.

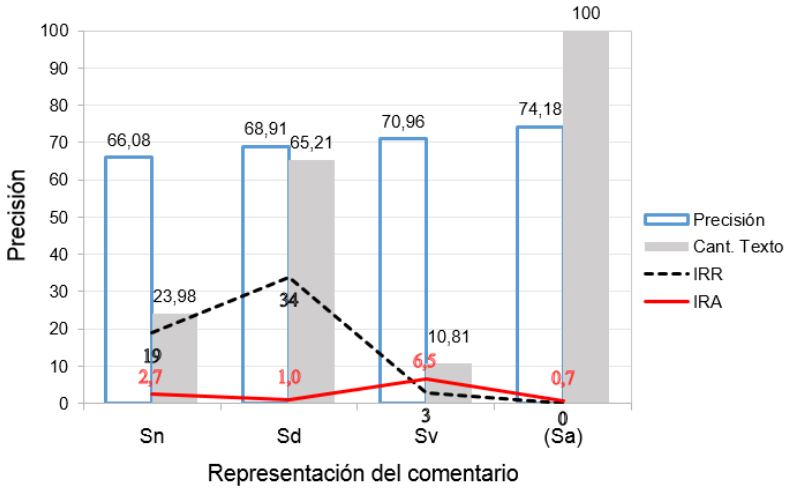


Figura 16: Relación entre precisión e índice de rendimiento (`IRA` y `IRR`) obtenidos con las cuatro representaciones de los comentarios: Sn, Sd, Sv y Sa.

Finalmente, otra ventaja de usar el segmento valorativo (S_v) en lugar de todo el comentario es que la bolsa de palabras no aumenta excesivamente con cada nuevo comentario que se incorpora al conjunto total de datos. La Figura 17 demuestra que cuando la representación empleada es S_v la curva de crecimiento, encargada de medir el número de palabras que componen el conjunto de entrenamiento ($eje - y$), no incrementa significativamente a medida que se agregan nuevos textos a dicho conjunto ($eje - x$). Por el contrario, la misma curva incrementa significativamente cuando está elaborada a partir de S_a, es decir, el comentario completo. Los datos de la figura se basan en un incremento hipotético del 5 y el 10 por ciento en el número de palabras para S_a y del 10 por ciento para S_v. Como queda representado en la figura, el crecimiento exponencial en el número de palabras para S_v siempre se mantiene por debajo del de S_a. Por ejemplo, cuando el número de comentarios es 150, el número de palabras para S_v es inferior a 1000 pero cercano a 5.000 para S_a; con 1.500 comentarios S_v se sigue manteniendo por debajo de las 1.000 palabras, mientras que S_a supera las 5.000 con un incremento teórico del 5 % y está próximo a las 15.000 palabras con un incremento, también teórico, del 10 %. Como resultado, el uso del segmento valorativo (S_v) comporta una reducción significativa en la memoria y el tiempo de procesamiento.

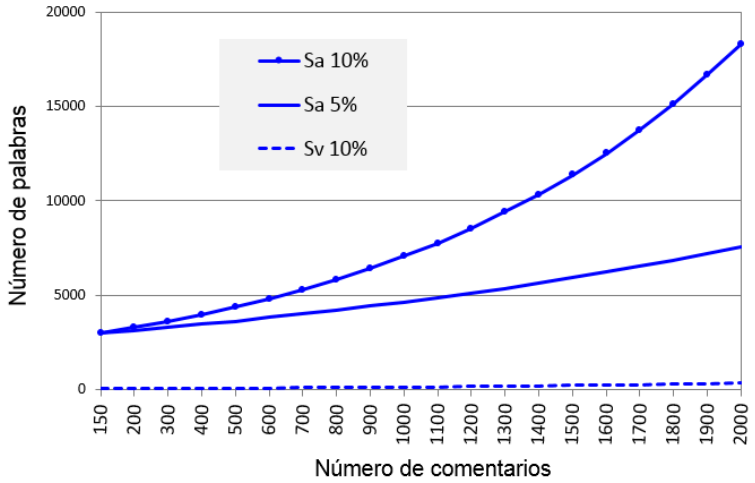


Figura 17: Aumento progresivo de la complejidad en las representaciones Sa y Sv debida al incremento del número de palabras tras la incorporación de nuevos comentarios al conjunto de entrenamiento.

5.3.2 Cálculo de la polaridad de comentarios en inglés

Para el cálculo de la polaridad en inglés he empleado los 90 comentarios obtenidos de MDTOD (45 comentarios con polaridad positiva y 45 con polaridad negativa). En este análisis, además de emplear una aproximación supervisada basada en BoW usando Weka [Witten y col., 1999], para clasificar los comentarios se evalúa el rendimiento de una aproximación no supervisada basada en léxicos (SO-CAL)[3]. Puesto que en el experimento anterior, el segmento valorativo demostró ser el más efectivo para calcular la polaridad del comentario, en este experimento aplico dos únicas representaciones

3 El sistema SO-CAL (*Semantic Orientation CALculator system* [Taboada, Brooke, Tofiloski y col., 2011] se describe en la Sección 3.4.2.

de los comentarios (segmento valorativo «S_v» versus comentario completo «S_a») usando como esquema de pesado un modelo binario. Esto se debe a que S_v/bin ha sido la mejor configuración detectada en la sección anterior para el análisis de la polaridad de los comentarios. Ambas representaciones, S_v y S_a, son usadas como entrada al sistema SO-CAL que evalúa su rendimiento, también, en términos de precisión.

Los resultados de este experimento aparecen en el Cuadro 40. Si comparamos los niveles de precisión obtenidos en este experimento con los conseguidos al calcular la polaridad de comentarios sobre hoteles en español, los resultados indican un incremento del rendimiento de los clasificadores al cambiar de dominio e idioma. De las dos aproximaciones usadas para el dominio de los coches en inglés, SO-CAL-S_a obtuvo la precisión más alta (87.9 %). Esto se debe a que, como se mencionó en el apartado 3.4.2, SO-CAL es una herramienta fundamentada en información lingüística[4] y diseñada específicamente para calcular la polaridad de los textos, de ahí que su desempeño supere al del clásico BoW.

Aproximaciones	SO-CAL		BoW+Weka	
Representaciones	S_a	S_v	S_a	S_v
Precisión	**87.9 %**	↓72.2 %	78.4 %	↑79.3 %
Cantidad de texto	100 %	18 %	100 %	18 %
IRA	0.8	4.0	0.7	4.4
IRR	0	28	95	15

Cuadro 40: Precisión obtenida al clasificar la polaridad de los comentarios sobre coches en inglés mediante dos aproximaciones (SO-CAL versus BoW) y dos representaciones diferentes (comentario completo versus segmento valorativo).

Adicionalmente, he de destacar que aunque el segmento valorativo S_v incrementa el rendimiento de los clasificadores de la polaridad bajo BoW (S_a = 78.4 % y S_v = 79.3 %), lo reduce bajo SO-CAL (S_a = 87.9 % y S_v = 72.2 %). En otras palabras, la omisión de los segmentos narrativo y descriptivo afecta negativamente el cálculo de la polaridad bajo una aproximación que tiene en cuenta la polaridad de base y la contextual de las palabras[5]. De este resultado se puede deducir que la polaridad en los segmentos narrativo y descriptivo se está expresando mediante mecanismos lingüísticos diferentes (posiblemente, más sofisticados) que la simple presencia/ausencia de determinados términos. El mismo resultado también me permite concluir que el segmento valorativo no está bien representado por los léxicos de polaridad. Por ejemplo, BoW asigna polaridad negativa a un comentario al detectar la presencia de la palabra *lemon* en su segmento valorativo[6]: S_v = *Maybe I just got a lemon*. Dado que (*i*) los diccionarios del SO-CAL no recogen la polaridad (negativa) de este término y (*ii*) S_v no contiene términos de polaridad, el SO-CAL es incapaz de asignar la polaridad adecuada al comentario usando sólo S_v.

En lo que respecta al índice de rendimiento (relación cantidad de texto y precisión), a pesar de que los segmentos valorativos en el corpus MDTOD

4 El SO-CAL ha sido diseñado para emplear las reglas de contexto inspiradas en la propuesta de Polanyi y Zaenen [2005] y los «modificadores contextuales de la valencia/polaridad» (*Contextual Valence Shifters*).

5 El SO-CAL aplica léxicos de polaridad para determinar la polaridad de base de las palabras y una serie de reglas para inferir su polaridad contextual.

6 Un número significativo de comentarios utilizan la palabra *lemon* («limón») para valorar negativamente un coche.

contienen un 60 % más de texto que los de HOpinion ($S_v(\text{HOpinion}) = 10.8\,\%$ y $S_v(\text{MDTOD}) = 18\,\%$), los resultados siguen favoreciendo el uso de la representación de los comentarios basada en los segmentos valorativos. Concretamente, el `IRA` más elevado se consigue con BoW-S_v=4.4 y el más bajo BoW-S_a=0.7. Este último `IRA` es el mismo que el obtenido con S_a para los comentarios sobre hoteles en español. En cuanto al `IRR`, que se calcula a partir de la precisión más elevada o de referencia (en este caso S_a en SO-CAL), el mejor resultado es también para BoW-S_v=15, mientras que el peor sigue siendo la representación que usa el comentario completo (BoW-S_a=95).

El resultado de los experimentos descritos en este apartado me ha permitido llegar a varias conclusiones. La primera de ellas es que el segmento valorativo es el más efectivo para detectar la polaridad de los comentarios bajo una representación supervisada del tipo BoW-S_v. Este resultado se presenta en términos de la relación entre precisión y la cantidad de texto (número de palabras) que es necesaria para generar cada representación S_x. En español, la precisión promedio de todos los clasificadores que usan BoW-S_v para obtener la polaridad del comentario es de 70 %, un valor que además de estar muy próximo al 71 % de BoW-S_a, sólo utiliza el 10.8 % del texto del comentario. En inglés, la precisión para BoW-S_v (79 %) llega a superar la de BoW-S_a (78 %), empleando tan sólo un 18 % del texto del comentario. No obstante, el análisis de comentarios en inglés también ha servido para concluir que el segmento valorativo no muestra el mismo buen desempeño bajo una aproximación no supervisada basada en el uso de léxicos (SO-CAL-S_v). Por lo tanto, en las siguientes secciones propongo mecanismos alternativos para identificar y analizar, exclusivamente, los segmentos narrativo y descriptivo de los comentarios sobre productos. El objetivo, en ambos casos, es determinar la función que cumple cada uno de estos segmentos en la expresión de la polaridad. En la Sección 5.6 analizo cuál es la función de cada segmento en la expresión de la polaridad a partir de los resultados conseguidos mediante los experimentos.

5.4 LAS SECUENCIAS NARRATIVAS APLICADAS AL ANÁLISIS DE LA POLARIDAD

Los experimentos presentados en la Sección 5.3 revelan que la omisión de los segmentos narrativo y descriptivo afecta negativamente el cálculo de la polaridad bajo un modelo léxico (SO-CAL-S_v). El objetivo de esta sección es evaluar un modelo alternativo para la identificación del segmento narrativo como paso previo al análisis de la polaridad de los comentarios. Para cumplir con este objetivo, en este apartado propongo una definición más elaborada del segmento narrativo según la cual, el segmento narrativo es aquel que está formado por el conjunto de las oraciones que en un comentario «narran» eventos relacionados temporalmente, es decir, por «secuencias narrativas» («*pasamos* a la habitación después de *haber hecho* el check-in»). Los experimentos que presento en los siguientes apartados están orientados a recuperar las secuencias narrativas de los comentarios y aplicarlas al cálculo de la polaridad. Concretamente, en el Apartado 5.4.1 describo algunos de los modelos que desde el Procesamiento del Lenguaje Natural (PLN) se han usado para identificar las secuencias narrativas de los textos. En el Apartado 5.4.2 presento una adaptación del modelo de Chambers y Jurafsky [2010] para identificar las secuencias narrativas de los comentarios. Y, en el Apartado

5.4.3 presento los experimentos en los que aplico el segmento narrativo al cálculo de la polaridad.

5.4.1 Las secuencias narrativas

El PLN y la IA se han interesado en la modelización de las secuencias temporales prototípicas de determinadas situaciones. Estas disciplinas han empleado estructuras de alto nivel como los guiones (*scripts*) y marcos (*frames*), desarrollados por Schank y Abelson [1975] y Minsky [1975] en los años setenta, para representar los eventos, sus relaciones causales y sus participantes. El ejemplo clásico de un guión es la secuencia de eventos que tienen lugar cuando una persona va a un restaurante y «se sienta», «le traen el menú», «ordena», «pide la cuenta», etc. La descripción de estas secuencias supone la selección de los eventos que forman parte de la narración y la organización temporal de dichos eventos. La selección de eventos o aprendizaje de la estructura narrativa de los eventos (*learning narrative event structure*) [Bejan, 2009] tiene como objetivo determinar el conjunto de eventos y entidades que integran situaciones (narrativas) prototípicas, por ejemplo, un proceso penal o judicial: eventos («condenar», «buscar», «declarar») y entidades («policía», «sospechoso», «jurado»). Por su parte, la organización temporal de eventos o aprendizaje del orden temporal de los eventos (*learning the temporal ordering of events*) [Barzilay y Lapata, 2005] busca determinar el orden temporal en que se suceden dichos eventos. Por ejemplo, un análisis del orden temporal determina que el «arresto» precede al «juicio» y que éste, a su vez, va antes de la «condena».

Los trabajos más recientes en comprensión del lenguaje natural aplican técnicas estadísticas para extraer, de manera automática, este conocimiento de los textos. Los métodos de aprendizaje supervisado y no supervisado que integran recursos como corpus anotados manualmente, son aplicados para representar eventos y otro tipo de conocimiento similar basado en el «sentido común» (*common-sense knowledge*). A continuación, comento algunos de los trabajos más representativos orientados al tratamiento computacional de los eventos y las narraciones.

Algunos trabajos representativos del tratamiento computacional de los eventos y las narraciones corresponden a Regneri y col. [2010], Hajishirzi, Hockenmaier y col. [2011], B. Li, Lee-Urban y col. [2012] y Chambers y Jurafsky [2009]. En primer lugar, Regneri y col. [2010] aplican aprendizaje no supervisado para detectar las frases que describen un mismo evento («sentarse a la mesa», «tomar asiento», etc.)[7] y el orden en que suelen presentarse en una narración (`script`). El procedimiento se basa en el uso de una matriz de frases semánticamente relacionadas donde se aplica el alineamiento múltiple de secuencias (*Multiple Sequence Alignment*, MSA). Sobre esta representación matricial se construye un «grafo temporal» que, mediante un algoritmo de agrupamiento, determina el orden en que se han de presentar los eventos. De manera similar, Hajishirzi, Hockenmaier y col. [2011] y Hajishirzi y Mueller [2012] analizan la forma de interpretar oraciones narrativas mediante la representación simbólica de los eventos, estados y entidades que ellas contienen. Su aproximación se vale de dos tipos de conocimiento: la descripción de los eventos y las entidades más importantes del dominio. Para la descripción temporal de los eventos (y estados) utilizan un lenguaje simbólico caracterizado por la presencia de condiciones y consecuencias:

7 Aunque parte de su trabajo consiste en detectar las diferentes realizaciones lingüísticas de un mismo evento, Regneri y col. [2010] aclaran que no se trata de un problema de paráfrasis.

$\langle evento(\vec{x}), condición(\vec{x}), consecuencia(\vec{x})\rangle$[8]. Por su parte, B. Li, Lee-Urban y col. [2012] y B. Li, Appling y col. [2012] exponen una técnica para identificar los eventos característicos de una determinada situación y su disposición temporal más habitual. Primero, los autores obtienen ejemplos reales de secuencias narrativas en diferentes ámbitos. Posteriormente, agrupan las oraciones que están semánticamente relacionadas (ej. «la policía detiene al criminal», «los agentes arrestan al ladrón») aplicando la similitud de cosenos. Por último, establecen restricciones entre eventos (precedencia, opcionalidad, exclusión) mediante un análisis de su frecuencia y probabilidad de aparición.

También Chambers y Jurafsky han estudiado el tratamiento computacional de la temporalidad: [Chambers, 2011], [Chambers y Jurafsky, 2010], [Chambers y Jurafsky, 2009], [Chambers y Jurafsky, 2008b] y [Chambers y Jurafsky, 2008a]. En su tesis doctoral, Chambers [2011] presenta un modelo para (*i*) identificar los eventos y los participantes (*entities*) que forman parte de una narración y (*ii*) inferir el orden de los eventos que allí aparecen. Su modelo parte del concepto de «cadena narrativa» (*narrative chain*). Según el autor, una cadena narrativa es un conjunto de eventos (parcialmente) ordenados que giran alrededor de una entidad común denominada «protagonista». La principal característica de las cadenas narrativas es que relacionan una única entidad o participante con una secuencia de eventos. A su vez, las cadenas narrativas se pueden extender para incorporar otros eventos y otros protagonistas dando lugar a «esquemas narrativos» (*narrative schemas*). Por ejemplo, en la Figura 18 aparecen dos cadenas narrativas que forman parte de un esquema narrativo relacionado con el procesamiento penal de un delincuente (Figura 18(a)).

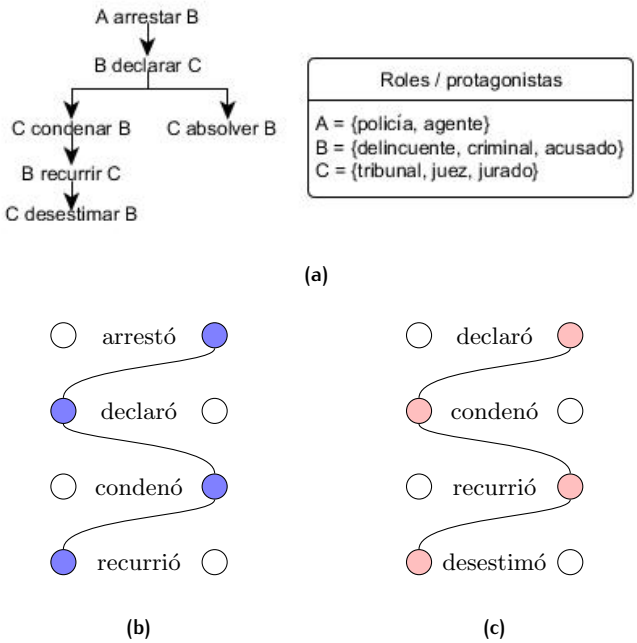


Figura 18: Ejemplo de un fragmento de «esquema narrativo» **(a)** compuesto por dos «cadenas narrativas» asociadas a los protagonistas: «delincuente» **(b)** y «tribunal» **(c)**. El ejemplo está basado en Chambers [2011, p. 18].

8 Por ejemplo, decir que «John cogió el ascensor» implica (condición) que antes «John entró en el edificio».

Siguiendo la notación gráfica utilizada por Chambers y Jurafsky, los círculos representan los diferentes protagonistas (roles) que participan en el esquema narrativo y que en el transcurso de la narración pueden aparecer designados bajo expresiones lingüísticas correferentes (ej. «policía», «agente»). La primera cadena (Figura 18(b)) tiene como protagonista al «delincuente» y la segunda (Figura 18(c)) al «tribunal». Por lo tanto, algunos de los eventos que se describen en el esquema son: «el policía arrestó al delincuente», «el delincuente declaró ante el tribunal», «el tribunal condenó al delincuente», etc.

Aunque, por definición, las cadenas narrativas presentan un orden parcial o natural de los eventos que contienen, Chambers [2011] aplica un método para establecer el orden temporal. El autor establece relaciones entre pares de eventos basándose en la frecuencia con la que comparten determinados argumentos gramaticales. Posteriormente, calcula la información mutua (*Pointwise Mutual Information*, PMI) entre pares de predicados con sujetos correferentes para inferir el conjunto de eventos que forman parte de una misma cadena narrativa. Por último, ordenan los eventos según la relaciones before/other (ej. ⟨arrestar BEFORE condenar⟩) valiéndose de un conjunto de rasgos temporales asociados a los verbos como son tiempo, aspecto, modalidad, relaciones de dependencia, *Part of Speech*, etc.

Además de su sencillez conceptual y elegancia metodológica, el modelo de Chambers se caracteriza por no requerir de un corpus anotado manualmente como paso previo para la identificación de las cadenas narrativas. Éstas características hacen de éste modelo una de las opciones más apropiadas para identificar las secuencias narrativas en los comentarios. En el apartado siguiente aplicamos las «cadenas narrativas» de Chambers y Jurafsky a la detección de las secuencias narrativas en los comentarios sobre productos como paso previo al tratamiento de su polaridad.

5.4.2 Clasificación automática de las secuencias narrativas

En esta sección aplico una versión simplificada y adaptada del modelo de Chambers y Jurafsky a la detección las secuencias narrativas que aparecen en los comentarios sobre productos. Considero que el análisis de la polaridad a partir de este tipo de segmentos no exige inferir el orden temporal de los eventos por lo que el trabajo está enfocado a la detección de las entidades y los eventos presentes en las cadenas narrativas.

Comienzo por definir un comentario (C) como un texto de opinión compuesto por un conjunto de oraciones narrativas (O_N), descriptivas (O_D) y valorativas (O_V), es decir, $C = \{O_N + O_D + O_V\} = \{o_1, o_2, o_3, ..., o_n\}$. Adicionalmente, una secuencia narrativa (O_{eN}) es un subconjunto de las oraciones que en O_N relatan eventos relacionados semántica y temporalmente: $O_{eN} \subseteq O_N$. Cada evento (e) es una tupla conformada por el verbo y sus argumentos: $e = \langle v, arg \rangle$ donde $arg \in \{suj, obj, prep\}$. Un segmento narrativo (S_N) se compone de todas las secuencias narrativas que forman parte del comentario: $S_N = \forall O_{eN} \in C$. Un caso concreto de S_N lo podemos ver en el ejemplo (11):

(11) *My dad bought me a Saturn for my graduation in 1992 before all the marketing hype (how embarrassing to be constantly asked if I went to Tennessee!). Shortly thereafter the problems started. ...*
Mi papá me compró un Saturn para mi graduación en 1992 antes de todo el boom publicitario (qué vergüenza que te pregunten constantemente si fuiste a Tennessee!). Poco después empezaron los problemas. ...

Con el fin de evidenciar la relación semántica entre los eventos de un comentario recurriré a la «presunción de la coherencia narrativa»[9] de Chambers y Jurafsky [2008b]. Según estos autores, **los verbos que comparten argumentos correferentes están relacionados semánticamente en virtud de la estructura narrativa del discurso**. Por ejemplo, en el fragmento (11) los verbos *bought* y *went* tienen pronombres correferentes (*me* e *I*), por lo que se establece una relación semántica entre ambos verbos: $\langle \texttt{bought}, \texttt{X}_{\texttt{objeto}} \rangle$ y $\langle \textit{went}, \texttt{X}_{\texttt{sujeto}} \rangle$. Esta relación queda representada en la Figura 19, en donde los círculos sombreados representan el elemento correferente *X* (en un caso *X* es objeto y en otro sujeto oracional). Los círculos en blanco son las entidades que no tienen ningún nexo correferencial: *my dad* y *Tennessee*.

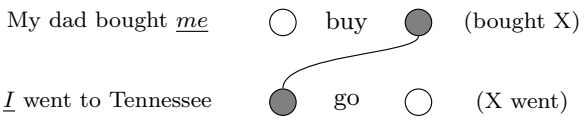


Figura 19: Relación semántica entre eventos (modelo básico).

El problema de esta representación es que solo captan 2 de los 6 eventos subrayados en el ejemplo (11). Para incluir los 4 eventos restantes se han de considerar, además de los verbos, las nominalizaciones deverbales (*graduation* y *hype*), las expresiones temporales (*before* y *shortly thereafter*) y los nexos coordinantes y subordinantes (*if*). Incluyo las nominalizaciones en la nueva representación puesto que estas construcciones lingüísticas se caracterizan por heredar la estructura argumental de los verbos de los que derivan; las expresiones temporales, por su parte, indican la duración o localización de un evento en relación con otro u otros eventos; de forma similar, los nexos coordinantes y subordinantes relacionan eventos aunque, en este caso, por vía gramatical. El resultado del nuevo análisis se puede ver en la Figura 20.

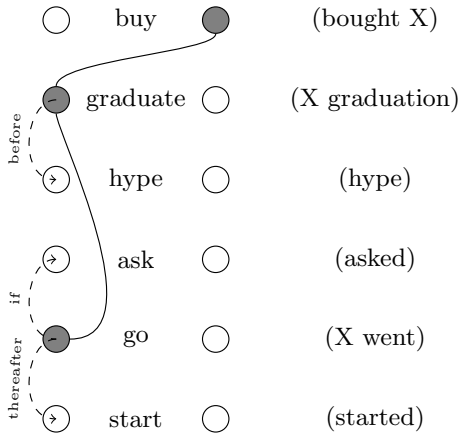


Figura 20: Relación semántica entre eventos (modelo extendido).

Como queda evidenciado en la Figura 20, el sustantivo deverbal *graduation* se incorpora de forma directa (comparte el argumento X) a la narración gracias al pronombre posesivo *my*. Por su parte los eventos *hype* y *ask* son

9 *Narrative Coherence Assumption.*

incorporados de forma indirecta (no comparten argumentos) por una expresión temporal (*before*) y por un nexo subordinante (*if*). Finalmente, el verbo *start* lo incluimos por la proximidad de ésta oración con las anteriores y porque está encabezada por una locución adverbial (*shortly thereafter*) que expresa un orden temporal.

El Algoritmo 1 describe el procedimiento que aplico para extraer de los comentarios las oraciones que conformarán las secuencias narrativas según el modelo expuesto en la Figura 20. La entrada del algoritmo es el comentario completo (C), concretamente, las oraciones que lo componen. En el Cuadro 41 presento el comentario del ejemplo (11) cuyas oraciones $C = \{o_1, o_2, o_3, o_4, o_5\}$ constituyen la entrada del Algoritmo 1. Las oraciones en negrita (o_1 y o_2) conforman el segmento narrativo del comentario que se pretende recuperar.

Algorithm 1 Detección automática de secuencias narrativas

Require: $C = \{o_1, o_2, o_3, ..., o_n\}$
1: $e_{list} \leftarrow \varnothing$ // lista de eventos
2: $O_{eN} \leftarrow \varnothing$ // secuencia narrativa
3: $primero \leftarrow false$ // variable booleana auxiliar
4: $segundo \leftarrow false$ // variable booleana auxiliar
5: $o_x \leftarrow \varnothing$ // oración auxiliar
6: $o_y \leftarrow \varnothing$ // oración auxiliar
7: **while** $\forall o_i \in C$ **do**
8: $deps = \{par_k : 1 \leqslant k \leqslant n_k\}$ donde $par_k = \langle núcleo, modificador \rangle$ // se obtienen las dependencias
9: $crefs = \{c_m : 1 \leqslant m \leqslant n_m\}$ donde $c_m = \langle palabra \rangle$ // se resuelven las correferencias
10: $nomin = \{n_s : 1 \leqslant s \leqslant n_s\}$ donde $n_s = \langle palabra \rangle$ // se identifican las nominalizaciones
11: $time = \{t_t : 1 \leqslant t \leqslant n_t\}$ donde $t_t = \langle palabra \rangle$ // se identifican las expresiones temporales
12: **for all** $par = \langle núcleo, modificador \rangle \in deps$ **do**
13: **if** $((núcleo = verbo | núcleo = nomin)$ **and** $(modificador = crefs | modificador = time))$ **then**
14: $e_{list} \leftarrow e_{list} \cup \{núcleo\}$ // se guarda el evento
15: **end if**
16: **end for**
17: **if** $|e_{list}| \geqslant 2$ **then**
18: $O_{eN} \leftarrow O_{eN} \cup \{o_i\}$ // se agrega la oración a la secuencia narrativa
19: **if** $(i+1) \leqslant n$ **then**
20: $o_x \leftarrow o_{i+1}$
21: **end if**
22: **if** $(i+2) \leqslant n$ **then**
23: $o_y \leftarrow o_{i+2}$
24: **end if**
25: **end if**
26: **while** $\forall t_t \in time$ **and** $(o_x \neq \varnothing || o_y \neq \varnothing)$ **do**
27: **if** $t_t \subset o_x$ **then**
28: $O_{eN} \leftarrow O_{eN} \cup \{o_x\}$ // se agrega la oración a la secuencia narrativa
29: $o_x \leftarrow \varnothing$
30: **end if**
31: **if** $t_t \subset o_y$ **then**
32: $O_{eN} \leftarrow O_{eN} \cup \{o_y\}$ // se agrega la oración a la secuencia narrativa
33: $o_y \leftarrow \varnothing$
34: **end if**
35: **end while**
36: **end while**
Ensure: O_{eN}

A continuación describo detalladamente cómo trabaja el Algoritmo 1. En primer lugar, mediante una llamada al parser de la Universidad de Stan-

o_1: My dad bought me a Saturn for my graduation in 1992 before all the marketing hype (how embarrassing to be constantly asked if I went to Tennessee!).
o_2: Shortly thereafter the problems started.
o_3: With a little research online you can find plenty of evidence that Saturns have a history of excessive oil consumption.
o_4: My 92 SL2 is a total lemon - blown head gasket, six alternators and a plethora of other problems with less than 60,000.
o_5: The dealer has not been helpful - saying that a blown head gasket at 57,000 miles is not uncommon (well for Saturns maybe!).

Cuadro 41: Ejemplo de entrada del Algoritmo 1 donde se muestran todas las oraciones del comentario.

ford[10] el Algoritmo obtiene las dependencias sintácticas (línea 8) y resuelve las correferencias de cada oración (línea 9). En el Cuadro 42 se muestran las dependencias con los pares núcleo-modificador que se obtienen al procesar la oración o_2 del ejemplo (11). En el Cuadro 43 aparece el listado de las entidades que comparten un nexo correferencial en el comentario. En este último cuadro se especifica la oración (o) a la que pertenece la entidad, la posición (p) que ocupa según el número de palabras en o, el identificador de las entidades que son correferentes ($crefs_{id}$) y la lista de entidades ($crefs_{ent}$).

```
root(ROOT-0, started-5)
advmod(thereafter-2, Shortly-1)
advmod(started-5, thereafter-2)
det(problems-4, the-3)
nsubj(started-5, problems-4)
```

Cuadro 42: Resultado del Algoritmo 1 (línea 8) para la obtención de las dependencias sintácticas.

o	p	$crefs_{id}$	$crefs_{ent}$
1	1	1	my
1	4	1	I
1	8	1	my
1	25	1	I
4	1	1	my
3	13	2	saturn
5	23	2	saturn
4	3	3	sl2
4	7	3	lemon
4	11	4	gasket
4	17	4	plethora

Cuadro 43: Resultado del Algoritmo 1 (línea 9) para la obtención de las entidades que comparten un nexo correferencial.

En segundo lugar, el Algoritmo 1 identifica las nominalizaciones basándose en un listado de nombres deverbales[11] que he extraído de *NomBank* (línea 10 del Algoritmo 1). En el Cuadro 44 tenemos el listado de nominalizaciones presentes en el comentario de ejemplo. La primera columna corresponde a la oración (o) de la que procede la nominalización y la segunda, al número

10 http://nlp.stanford.edu/software/lex-parser.shtml
11 Considero sólo los sustantivos que se derivan directamente de verbos.

de palabra (p). En la tercera columna tenemos las nominalizaciones y en la cuarta los verbos de los que proceden según el *NomBank*. Posteriormente, en la línea 11 del Algoritmo 1, se aplica un procedimiento similar para identificar las expresiones temporales a partir de *TimeBank* [Pustejovsky y col., 2003] y de un listado de adverbios, adjetivos y preposiciones extraído del mismo corpus (ej. *after*, *immediately*, *follows*, *meanwhile*, etc.).

o	p	nominalización	verbo
1	9	graduation	graduate
1	15	marketing	market
1	16	hype	exaggerate
3	4	research	research
3	11	evidence	evidence
3	16	history	record
3	20	consumption	consume
4	10	head	head
5	2	dealer	deal
5	12	head	head

Cuadro 44: Resultado del Algoritmo 1 (línea 10) para la obtención de las nominalizaciones presentes en el comentario.

En tercer lugar, el Algoritmo 1 selecciona los pares núcleo - modificador (línea 12) y, si el núcleo es un verbo o una nominalización y el modificador un argumento correferente, o una expresión temporal, agrega el núcleo a la lista de eventos (línea 14 del Algoritmo 1). Dado que una secuencia narrativa se compone como mínimo de dos eventos, sólo la oración que contiene dos o más eventos (línea 17) pasa a formar parte de la secuencia narrativa (línea 18 del Algoritmo 1). Adicionalmente, siempre que sea posible, el Algoritmo 1 obtiene la referencia a las dos oraciones siguientes (líneas 19 a la 24) y si alguna de estas dos oraciones contiene una expresión temporal, también la incluye en la secuencia narrativa (líneas 26 a la 34).

En cuarto lugar, el Algoritmo 1 retorna el conjunto de oraciones O_{eN} seleccionadas que conforman el segmento narrativo S_N del comentario C. En la primera y quinta columnas del Cuadro 45 aparecen las oraciones y los eventos seleccionados por el algoritmo. En la tercera columna del mismo cuadro tenemos el identificador que vincula cada evento con su entidad correferente ($crefs_{id}$). En la cuarta columna aparece la expresión temporal ($time$) que se utiliza para seleccionar un evento cuando éste no comparte un argumento correferente. Finalmente, es necesario aclarar que aunque en la Figura 20 se incluye el evento *ask* como parte del modelo, la implementación actual del algoritmo no capta dicho evento. Una tarea pendiente, como trabajo futuro, es buscar alternativas al tratamiento de estos eventos «aislados».

o	p	$crefs_{id}$	$time$	eventos
1	3	1	∅	buy
1	9	1	∅	graduation
1	16	∅	before	hype
1	26	1	∅	go
2	5	∅	thereafter	start

Cuadro 45: Salida del Algoritmo 1 para la obtención de los eventos que forman parte del segmento narrativo del comentario del Cuadro 41.

5.4.3 Cálculo de la polaridad mediante el segmento narrativo

Al principio de esta sección he asumido que el segmento narrativo está constituido por el conjunto de las oraciones que en un comentario forman parte de una secuencia narrativa. A continuación, presento los experimentos que he realizado con el fin de determinar el rendimiento que exhibe el segmento narrativo para el cálculo de la polaridad de los comentarios sobre productos. Para realizar estos experimentos he usado los 2.547 comentarios sobre coches, hoteles y auriculares que forman parte del corpus de opiniones en inglés de Cruz Mata [2012]. El procedimiento que aplico para el análisis es el siguiente. En primer lugar, identifico las secuencias narrativas de forma automática usando, en este caso, el Algoritmo 1 implementado en la Sección 5.4.2. En segundo lugar, calculo la polaridad de los comentarios contrastando diferentes configuraciones de datos.

5.4.3.1 *Configuraciones*

El análisis de la polaridad de los comentarios se ha realizado considerando seis configuraciones basadas en la supresión de diferentes fragmentos del texto: una configuración de referencia que usa el texto completo, tres configuraciones objetivo que eliminan partes del texto de manera controlada y dos configuraciones de control que eliminan partes del comentario de forma aleatoria. A continuación describo y ejemplifico la forma que tienen tales configuraciones.

CONFIGURACIÓN 1 (C_{ref}): Es la configuración de referencia. Bajo esta configuración utilizo el comentario en su totalidad (100 % del texto) para el cálculo de la polaridad. Si tomo como ejemplo la versión completa del comentario (11), C_{ref} está compuesto por cinco oraciones $C_{ref} = \{o_1, o_2, o_3, o_4, o_5\}$ y 98 palabras, tal y como se muestra a continuación:

1 My$_1$ dad$_2$ bought$_3$ me$_4$ a$_5$ Saturn$_6$ for$_7$ my$_8$ graduation$_9$ in$_{10}$ 1992$_{11}$ before$_{12}$ all$_{13}$ the$_{14}$ marketing$_{15}$ hype$_{16}$ (how$_{17}$ embarrassing$_{18}$ to$_{19}$ be$_{20}$ constantly$_{21}$ asked$_{22}$ if$_{23}$ I$_{24}$ went$_{25}$ to$_{26}$ Tennessee$_{27}$!).
2 Shortly$_{28}$ thereafter$_{29}$ the$_{30}$ problems$_{31}$ started$_{32}$.
3 With$_{33}$ a$_{34}$ little$_{35}$ research$_{36}$ online$_{37}$ you$_{38}$ can$_{39}$ find$_{40}$ plenty$_{41}$ of$_{42}$ evidence$_{43}$ that$_{44}$ Saturns$_{45}$ have$_{46}$a$_{47}$ history$_{48}$ of$_{49}$ excessive$_{50}$ oil$_{51}$ consumption$_{52}$.
4 My$_{53}$ 92$_{54}$ SL2$_{55}$ is$_{56}$ a$_{57}$ total$_{58}$ lemon$_{59}$ -$_{60}$ blown$_{61}$ head$_{62}$ gasket$_{63}$, six$_{64}$ alternators$_{65}$ and$_{66}$ a$_{67}$ plethora$_{68}$ of$_{69}$ other$_{70}$ problems$_{71}$ with$_{72}$ less$_{73}$ than$_{74}$ 60,000$_{75}$.
5 The$_{76}$ dealer$_{77}$ has$_{78}$ not$_{79}$ been$_{80}$ helpful$_{81}$ -$_{82}$ saying$_{83}$ that$_{84}$ a$_{85}$ blown$_{86}$ head$_{87}$ gasket$_{88}$ at$_{89}$ 57,000$_{90}$ miles$_{91}$ is$_{92}$ not$_{93}$ uncommon$_{94}$ (well$_{95}$ for$_{96}$ Saturns$_{97}$ maybe$_{98}$!).

CONFIGURACIÓN 2 (nar_{ext}): Es una de las configuraciones objetivo. En nar_{ext} el análisis de la polaridad se realiza usando únicamente los segmentos narrativos (S_N), por lo que nar_{ext} informa sobre el beneficio que comporta utilizar este tipo de segmento para el cálculo de la polaridad. Esta configuración detecta las secuencias narrativas (O_{eN}) que conforman los comentarios sobre productos a partir del modelo extendido (ver Figura 20). Es importante aclarar que el modelo extendido clasifica y selecciona como narrativas un número mayor de oraciones puesto que, además de los verbos, utiliza las nominalizaciones, las expresiones temporales y los nexos coordinantes y subordinantes. Nuevamente, si tomo como ejemplo el comentario (11), $nar_{ext} = \{o_1, o_2\}$:

```
1 My dad bought me a Saturn for my graduation in 1992 before all
      the marketing hype (how embarrassing to be constantly asked
      if I went to Tennessee!).
2 Shortly thereafter the problems started.
```

CONFIGURACIÓN 3 (des_{bas}): Es la segunda configuración objetivo. Con des_{bas} el análisis de la polaridad se ejecuta usando los segmentos no narrativos, esto es, la parte del comentario que queda tras eliminar las secuencias narrativas (O_{eN}). Por comodidad me referiré a los segmentos no narrativos como segmentos descriptivos ($des = O_D \cup O_V$). des_{bas} selecciona y elimina las secuencias narrativas a partir del modelo básico (ver Figura 19), es decir, el modelo que menos oraciones cataloga como narrativas. Por lo tanto, $des_{bas} = \{o_2, o_3, o_4, o_5\}$:

```
2 Shortly thereafter the problems started.
3 With a little research online you can find plenty of evidence
      that Saturns have a history of excessive oil consumption.
4 My 92 SL2 is a total lemon – blown head gasket, six alternators
      and a plethora of other problems with less than 60,000.
5 The dealer has not been helpful – saying that a blown head gasket
      at 57,000 miles is not uncommon (well for Saturns maybe!).
```

CONFIGURACIÓN 4 (des_{ext}): Es la tercera configuración objetivo. Al igual que des_{bas}, des_{ext} utiliza los segmentos no narrativos para efectuar el análisis de la polaridad pero, en este caso, aplicando el modelo extendido de la Figura 20. Las configuraciones objetivo des_{bas} y des_{ext} informan directamente sobre el impacto que tiene la omisión de las secuencias narrativas en el cálculo de la polaridad. Dado que el modelo extendido para la detección de secuencias narrativas reduce el número de oraciones clasificadas como «descriptivas», entonces, $des_{ext} = \{o_3, o_4, o_5\}$:

```
3 With a little research online you can find plenty of evidence
      that Saturns have a history of excessive oil consumption.
4 My 92 SL2 is a total lemon – blown head gasket, six alternators
      and a plethora of other problems with less than 60,000.
5 The dealer has not been helpful – saying that a blown head gasket
      at 57,000 miles is not uncommon (well for Saturns maybe!).
```

CONFIGURACIÓN 5 (ctr_{40}): Es una de las configuraciones de control. En ctr_{40}, el análisis de la polaridad se realiza usando un 40 % del comentario. Esta configuración de control selecciona un número similar de palabras a la configuración objetivo des_{bas}, basada en la omisión de los segmentos narrativos. A continuación se muestra el comentario (11) tras eliminar el 60 % del texto:

```
1 the problems started. With a little research online you can find
      plenty of evidence that Saturns have a history of excessive
      oil consumption. My 92 SL2 is a total lemon – blown head
      gasket, six alternators and a plethora of other problems
```

CONFIGURACIÓN 6 (ctr_{30}): Es la segunda configuración de control. Con ctr_{30} el análisis de la polaridad se realiza usando solo un 30 % del comentario, seleccionado aleatoriamente. Esta configuración de control

selecciona un número similar de palabras a la configuración objetivo des_{ext}, basada en la omisión de los segmentos narrativos. A continuación se incluye el comentario (11) tras eliminar el 70 % del texto:

```
1 the problems started. With a little research online you can find
    plenty of evidence that Saturns have a history of excessive
    oil consumption. My 92 SL2 is a total lemon
```

5.4.3.2 *Resultados*

Con el fin de determinar la función que cumplen las secuencias narrativas en el cálculo de la polaridad de los comentarios sobre productos, además de las configuraciones definidas en la sección anterior (C_{ref}, nar_{ext}, des_{bas}, des_{ext}, ctr_{40} y ctr_{30}), he analizado los textos del corpus de Cruz Mata [2012] (MDTOD) teniendo en consideración el dominio (coches, auriculares y hoteles) y la polaridad (positiva + y negativa –) de los comentarios. El resumen de los datos empleados para este experimento aparece en el Cuadro 46.

Dominio	Muestras	Min. Pals.	Max. Pals.	Media Pals.	Total Pals.
coches	972	54	6462	508	466.542
auriculares	587	39	923	239	113.413
hoteles	988	34	4993	639	604.541

Cuadro 46: Resumen de los datos empleados para calcular la polaridad de los comentarios mediante los segmentos narrativos.

Los resultados del análisis se pueden ver en el Cuadro 47. En este cuadro presento los niveles de precisión que se obtienen al predecir la polaridad de los comentarios. Para el análisis de la polaridad he usado *Semantic Orientation CALculator* (SO-CAL) [Taboada, 2011]. SO-CAL utiliza diccionarios de palabras anotadas con su orientación semántica en una escala que va del 5 para los términos más positivos (*exquisite*) a -5 para los más negativos (*horrific*). SO-CAL también incorpora modificadores de la polaridad como son los intensificadores (***most*** *excellent*) y la negación (***not*** *good*).

El Cuadro 47 también muestra el índice de rendimiento absoluto (IRA), es decir, la relación entre la precisión y la cantidad de texto necesaria para obtener dicha precisión[12]. En la parte baja del cuadro aparecen los promedios de las precisiones (Promedios) agrupados según la polaridad de los comentarios: todos los comentarios (+-), comentarios positivos (+) y comentarios negativos (-). El IRA_{dif} (que explico más adelante) y la cantidad o longitud de texto usada en cada configuración (tanto en términos de porcentajes (%) como de número de palabras (pals.), cierran el Cuadro 47.

Las principales observaciones que extraigo de este análisis son las siguientes:

- C_{ref} obtuvo las mejores precisiones en el cálculo de la polaridad pero su rendimiento es muy bajo puesto que requiere de todo el texto de los comentarios (más de 1 millón de palabras) para alcanzar tales valores.

12 IRA aparece definido en la fórmula (3) de la página 82.

Dominio	pol.	C_{ref}	ctr_{40}	ctr_{30}	nar_{ext}	des_{bas}	des_{ext}
Coches	+	89.4	71.4	72.2	77.8	83.5	84.8
	IRA	0.8	1.7	2.4	1.2	1.8	2.1
	–	77.3	75.5	72.6	67.2	68.2	66.7
	IRA	0.7	1.8	2.4	1.1	1.5	1.6
Auriculares	+	82.9	65.8	66.6	69.7	82.5	81.7
	IRA	0.8	1.6	2.2	1.1	1.8	2.0
	–	72.4	68.1	68.1	60.3	65.7	62.4
	IRA	0.7	1.7	2.2	0.9	1.4	1.5
Hoteles	+	95.7	87.2	86.7	78	94.5	95.5
	IRA	0.9	2.1	2.8	1.2	2.1	2.4
	–	77.5	67	67.5	66.7	69	67.2
	IRA	0.7	1.6	2.2	1.1	1.5	1.7
Promedios	+-	82.5	72.5	72.2	69.9	77.2	76.3
	+	89.3	74.8	75.1	75.1	86.8	87.3
	-	75.7	70.2	69.4	64.7	67.6	65.4
IRA_{dif}		0.13	0.2	0.2	0.13	0.43	**0.56**
Longitud	%	100	40	30	60.6	44.2	39.4
	pals.	1.184.496	473.798	355.348	717.805	524.505	466.691

Cuadro 47: Resultados del cálculo de la polaridad en los comentarios.

La precisión más alta bajo esta configuración es de 95.7 % (hoteles) sin embargo en ningún momento el índice de rendimiento absoluto supera el valor del 1.0.

- ctr_{40} y ctr_{30} representan el caso contrario a C_{ref}: si bien el índice de rendimiento absoluto llega a superar el 2.0 (ctr_{40} / hoteles / + = 2.1 y ctr_{30} / hoteles / + = 2.8), sus niveles de precisión son más bajos que el de otras configuraciones (72 % de promedio, tanto en ctr_{40} como en ctr_{30}).

- nar_{ext} es la configuración menos eficaz: tiene los niveles de precisión más bajos de todas las configuraciones y el índice de rendimiento absoluto es modesto. Por ejemplo, la precisión para el dominio de los auriculares con polaridad negativa es de tan solo el 60.3 % con un rendimiento del 0.9.

- des_{bas} y des_{ext} presentan los niveles de precisión más próximos a C_{ref} en cuanto al cálculo de la polaridad positiva utilizando tan solo el 39.4 % del comentario (466.691 palabras) (ver celdas sombreadas en el Cuadro 47). Estos resultados contrastan significativamente con su bajo rendimiento en el cálculo de la polaridad negativa. El contraste al que hacemos referencia es más evidente en des_{ext} que en des_{bas}. Por ejemplo, en el dominio de los hoteles la precisión de aciertos positivos es del 95.5 % frente a un 67.2 % de precisión en aciertos negativos. De la misma forma, si atendemos a los valores del IRA en todas las configuraciones se observa que únicamente en des_{bas} y des_{ext} existe una divergencia significativa asociada a la polaridad de los comentarios. Esta divergencia la podemos concretar calculando el promedio de las diferencias del índice de rendimiento absoluto para cada representación (IRA_{dif}). En la fórmula (5) tenemos, a manera de ejemplo, el cálculo de IRA_{dif} para des_{ext}. Como se puede observar en el Cuadro 47, la

polaridad de los comentarios incide especialmente sobre las configuraciones des_{ext} (IRA_{dif}=0.56) y des_{bas} (IRA_{dif}=0.43). Por el contrario, el IRA_{dif} de las configuraciones restantes no superan el 0.2, es decir, que en ellas la incidencia de la polaridad del comentario es mínima.

$$\begin{aligned} \texttt{IRA}\ des_{ext}\ coches (2.1 - 1.6 = 0.5) \\ \texttt{IRA}\ des_{ext}\ auriculares (2.0 - 1.5 = 0.5) \\ \texttt{IRA}\ des_{ext}\ hoteles (2.4 - 1.7 = 0.7) \\ *** \\ \texttt{IRA}_{dif}\ des_{ext} = \frac{0.5 + 0.5 + 0.7}{3} = \mathbf{0.56} \end{aligned} \tag{5}$$

La conclusión general que se desprende del análisis del segmento narrativo es que existe una divergencia importante asociada al uso de este tipo de segmento y la polaridad que expresa. Los resultados indican que la narración no es muy útil para detectar la polaridad positiva pero sí la negativa: al omitir las secuencias narrativas de los comentarios (configuraciones des_{bas} y des_{ext}) estamos descartando información relevante para entender las opiniones negativas. Como se verá en la siguiente sección, esta divergencia no es exclusiva del segmento narrativo.

5.5 LA COMPLEJIDAD SINTÁCTICA APLICADA AL ANÁLISIS DE LA POLARIDAD

Los experimentos presentados en la Sección 5.3 revelan que la omisión de los segmentos narrativo y descriptivo afecta negativamente el cálculo de la polaridad bajo un modelo léxico (SO-CAL-S_v). El objetivo de esta sección es evaluar un modelo alternativo para la identificación del segmento descriptivo como paso previo al análisis de la polaridad de los comentarios. Para cumplir con este objetivo, en este apartado considero que el segmento descriptivo es aquel que está formado por el conjunto de las oraciones que en un comentario «describen» positiva o negativamente un producto o alguna de sus características («este coche consume demasiada gasolina»). Los experimentos que presento en los siguientes apartados están orientados a clasificar estas oraciones según su «complejidad sintáctica» y aplicarlas al cálculo de la polaridad. Concretamente, en el Apartado 5.5.1 defino la complejidad sintáctica. En el Apartado 5.5.2 uso diversas métricas de la complejidad sintáctica para clasificar de forma automática las oraciones que expresan la misma polaridad y la polaridad opuesta a la del comentario. Finalmente, en el Apartado 5.5.3, describo varios experimentos orientados a determinar el rendimiento que presentan estas oraciones en conjunto para el cálculo de la polaridad de los comentarios.

Dos hipótesis motivan el uso de la complejidad sintáctica al análisis de la polaridad:

HIPÓTESIS 1

Hipotetizo que las oraciones que expresan la misma polaridad que la del comentario (en adelante, oraciones simétricas) suelen ser sintácticamente menos complejas que las oraciones que expresan una polaridad opuesta (en adelante, oraciones asimétricas).

Partiendo de la Hipótesis 1 es posible formular la Hipótesis 2:

HIPÓTESIS 2

Hipotetizo que es posible mejorar los niveles de precisión en el cálculo de la polaridad omitiendo de los textos las oraciones asimétricas, es decir, las que expresan una polaridad opuesta a la del comentario completo.

Las oraciones descriptivas simétrica (12) y asimétrica (13) sirven para ejemplificar la hipótesis 1: en (12) aparece una oración extraída de un comentario positivo que se caracteriza por su escasa complejidad; en (13) se muestra una oración extraída de un comentario, también positivo, que se caracteriza por ser más compleja que la de (12). Por su parte, el comentario (14) ejemplifica la hipótesis 2. Este es un ejemplo de un comentario con polaridad positiva en el que se han identificado en negrita las oraciones que erróneamente inducen a asignar una polaridad negativa a todo el texto (oraciones descriptivas asimétricas) y en cursiva las oraciones que acertadamente inducen a asignar una polaridad positiva al texto (oraciones descriptivas simétricas). Las oraciones restantes no son descriptivas y no se tendrán en consideración en este análisis.

(12) **Oración simétrica:** Es un hotel muy acogedor.

(13) **Oración asimétrica:** A pesar de todo, creo que para ser un cinco estrellas es un hotel que deja mucho que desear.

(14) Estuvimos alojados en ese hotel durante tres días. **Es un hotel que a primera vista parece un poco viejo, anticuado y con una urgencia en cuanto al mantenimiento. Creo que para ser un cuatro estrellas es un hotel que deja mucho que desear en cuanto a su estética.** *A pesar de todo, las habitaciones están siempre limpias. El personal es agradable. El desayuno correcto. Muy bien ubicado.* En general es recomendable.

A continuación, en el apartado 5.5.1, defino en qué consiste la complejidad sintáctica y presento las métricas que empleo para calcularla. Posteriormente, en el apartado 5.5.2, presento un experimento orientado a analizar la validez de la Hipótesis 1, es decir, si es posible clasificar las oraciones simétricas y asimétricas mediante el cálculo de su la complejidad sintáctica. Mediante un segundo experimento, en el apartado 5.5.3, determino la validez de la Hipótesis 2. En este experimento evalúo si la omisión de las oraciones asimétricas de los comentarios mejora el rendimiento de un sistema para el cálculo de la polaridad.

5.5.1 La complejidad sintáctica

La complejidad sintáctica se define como «la gama de estructuras que surgen en la producción del lenguaje y el grado de sofisticación de tales estructuras» [Ortega, 2003]. A pesar de que no existe un consenso general entre los investigadores sobre cómo determinar la complejidad sintáctica, su cálculo suele depender del uso de oraciones o cláusulas incrustadas[13] (comparar las oraciones 15a y 15b) y el orden no canónico de las palabras (comparar las oraciones 16a y 16b). Para efectos de este análisis, definiré la complejidad sintáctica como la extensión y complicación de las unidades sintácticas del texto

13 Una cláusula incrustada (*embedding clause*) es aquella que se integra como constituyente de una cláusula regente.

a través de unos cuantos recursos sintácticos que favorecen la actualización y la producción de los significados.

(15) a. *I eat and you cook.*
b. *I eat* `if you cook`.

(16) a. *The student that met the teacher* (subject relative clause).
b. *The student that the teacher met* (direct object relative clause).

Gran parte de las métricas de la complejidad sintáctica que se han desarrollado están relacionadas con la adquisición de segundas lenguas, los trastornos del lenguaje o el envejecimiento cognitivo. En Jalilevand y Ebrahimipour [2014] se aplica el IPSyn (*Index of Productive Syntax*), el DSS (*Developmental Sentence Scoring*) y el MLU (*Mean Length of Utterance*), para contrastar las «habilidades lingüísticas» entre niños sanos y niños con alguna discapacidad en el uso del lenguaje. En Kemper y col. [2011] se estudia el procesamiento de estructuras sintácticas complejas en adultos mayores y jóvenes. Una síntesis de estas y otras métricas aparece en Moyle y S. Long [2013].

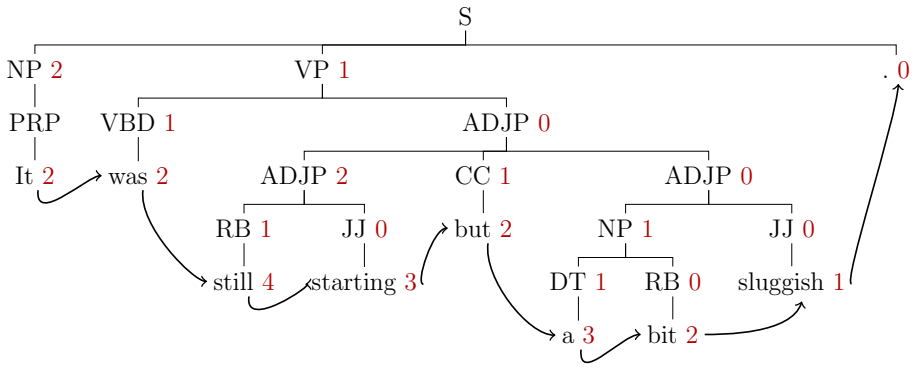
No obstante, el acto de formular una opinión es, ante todo, una actividad cognitiva que no depende del proceso de adquisición o desarrollo del lenguaje. Por este motivo, para calcular la complejidad sintáctica en los comentarios sobre productos, hemos seleccionado tres métricas que cuantifican el nivel de procesamiento cognitivo que se requiere para producir diferentes clases de construcciones sintácticas. Estas medidas corresponden a los índices de Yngve [Yngve, 1960], Frazier [Frazier y Clifton, 1998] y Pakhomov [Pakhomov y col., 2011]. Estas medidas tienen en común el que relacionan el tipo de oración con la capacidad de memoria requerida para su procesamiento.

5.5.1.1 *Índice de Yngve*

Yngve [1960] asume que existe una relación entre el coste de procesamiento asociado a la complejidad sintáctica y la capacidad individual de la memoria de trabajo o memoria operativa, es decir, la memoria que permite manipular y mantener temporalmente la información que se va necesitando durante la realización de cualquier tarea cognitiva. Yngve [1960] demostró que el número consecutivo de constituyentes ubicados a la izquierda en un árbol sintáctico determina la capacidad de almacenamiento temporal requerida para generar y procesar una oración. Según este modelo, las oraciones sintácticamente complejas son más difíciles de construir y comprender puesto que presentan una mayor «profundidad sintáctica» (*syntactic depth*), es decir, un número importante de constituyentes y mayor anidamiento de los mismos. Para Yngve [1960] la «profundidad sintáctica» de una palabra representa el número de constituyentes gramaticales planteados o proyectados que aún no han sido realizados durante la producción de una oración.

La profundidad de Yngve (*Yngve's depth algorithm*) se determina asignando un índice a cada uno de los nodos que forman parte de un constituyente en un árbol sintáctico, empezando por el nodo que está localizado más a la derecha (nodo con el índice 0). La profundidad de una palabra corresponde a la suma de todos los índices que conectan la palabra al nodo raíz (`S`). La Figura 21 ilustra el cálculo de la profundidad de Yngve en la oración *it was still starting but a bit sluggish*. Por ejemplo, la profundidad del verbo `was` es 2, puesto que `VBD = 1` y `VP = 1`. En la misma figura, `Total_Ydepth` es la suma

de la profundidad de todas las palabras y `Mean_Ydepth` es el total dividido por el número de palabras.



total_Ydepth: $2 + 2 + 4 + 3 + 2 + 3 + 2 + 1 + 0 = 19$
mean_Ydepth: 2.1

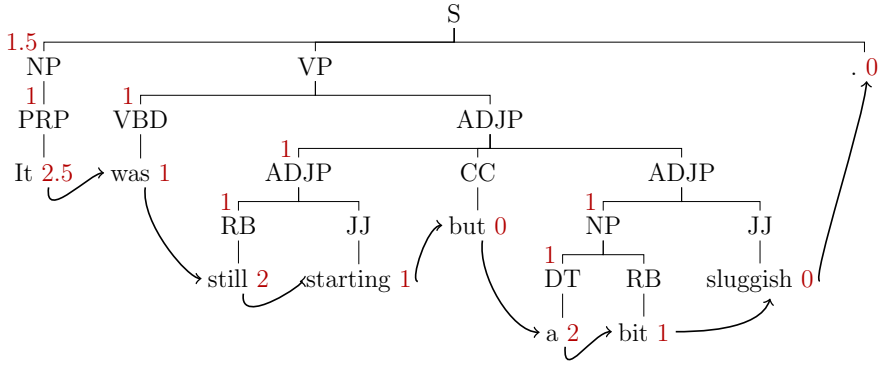
Figura 21: Cálculo del índice de profundidad de Yngve.

5.5.1.2 *Índice de Frazier*

Frazier [1985] y Frazier y Clifton [1998] postulan que la producción y comprensión del lenguaje están basados en un principio «universal» de economía cognitiva. Por lo tanto, ante dos estructuras sintácticas posibles, una simple y la otra compleja, los hablantes/oyentes tienden a elegir la simple. Un ejemplo de ello es la estrategia de «cierre tardío» o «adjunción baja». De acuerdo con Frazier, lo habitual es la adjunción de cada nuevo constituyente al sintagma procesado más recientemente: no se da comienzo a una cláusula o sintagma nuevo mientras se puedan seguir adjuntando componentes a la estructura que está siendo procesada en ese momento. Por lo tanto, en una oración con doble antecedente como «el policía entrevistó a la hija del coronel que tuvo un accidente», la estrategia de cierre tardío pronostica una tendencia a asignar la cláusula de relativo «que tuvo un accidente» al último sintagma nominal («el coronel») en vez de al primero («la hija»). El objetivo general de estas estrategias es respetar las limitaciones de capacidad de la memoria operativa para así aumentar la rapidez y eficiencia con la que la información nueva se va integrando en el análisis. El objetivo específico de estas estrategias es usar el menor número de nodos posibles para vincular cada nueva palabra en la estructura que se está construyendo [Carreiras, 1992].

El índice de Frazier [Frazier, 1985] se basa en el número de nodos no terminales que deben ser recorridos cuando se procesa una oración. Si bien, la complejidad de un constituyente puede determinarse simplemente dividiendo el número de nodos no terminales por el número de nodos terminales de la oración, Frazier [1985] aplica otro procedimiento para calcular la complejidad sintáctica de una oración. Dicho procedimiento consiste en asignar un valor fijo equivalente a 1 a los nodos no terminales que se localizan más a la izquierda de cada constituyente en un árbol sintáctico. El nodo que conecta directamente con el nodo raíz (S) le corresponde un valor especial de 1.5 ya que de esta manera las cláusulas embebidas contribuyen más al cálculo de la complejidad. El peso de una palabra corresponde a la suma de todos los valores (1s) que conectan la palabra en su recorrido hacia el nodo raíz (S).

La Figura 22 ilustra el cálculo del índice de Frazier en la oración *it was still starting but a bit sluggish*. Por ejemplo, el verbo `was` tiene un peso de 1 puesto que `VBD = 1` y `VP = ∅`. En la misma figura, `Total_Fdepth` es la suma de todos los índices de las palabras y `Mean_Fdepth` es el total dividido por el número de palabras.



total_Fdepth: $2.5 + 1 + 2 + 1 + 0 + 2 + 1 + 0 + 0 = 9.5$
mean_Fdepth: 1.05

Figura 22: Cálculo del índice de Frazier.

5.5.1.3 *Índice de Pakhomov*

El método que propone Pakhomov y col. [2011] para medir la complejidad sintáctica de una oración está inspirado en Gibson [1998] y Gibson [2000]. Gibson asume, en su teoría de la localidad de las dependencias sintácticas, que el sistema de procesamiento del lenguaje está limitado por el número de computaciones que puede realizar o por la cantidad finita de información que puede manipular o mantener activa mientras procesa. La complejidad de procesamiento de una estructura depende de la longitud de sus dependencias sintácticas. La teoría predice que las estructuras cuyas dependencias son de mayor longitud serán más difíciles de procesar que las de menor longitud debido al coste de almacenamiento y al coste de integración derivado de conectar sintácticamente palabras que mantienen relaciones de dependencia.

Basándose en esta teoría, Pakhomov y col. [2011] aplican el parser de la Universidad de Stanford para identificar las unidades léxicas que mantienen alguna relación de dependencia en el interior de una oración. En la aproximación de Pakhomov y col. [2011], cada relación de dependencia recibe una puntuación que mide la distancia que separa un núcleo de su modificador. Para ello, se calcula la diferencia entre las posiciones que ocupan el par de palabras en la oración. La Figura 23 ilustra el cálculo del índice de Pakhomov en la oración *it was still starting but a bit sluggish*. Por ejemplo, la distancia de los dos términos en la relación `nsubj` (sujeto nominal) es 3 puesto que $4_{\texttt{starting}} - 1_{\texttt{it}} = 3$. En la misma figura, `Total_SynDepLen` es la suma de todas las dependencias y `Mean_SynDepLen` es el total dividido por el número de dependencias.

En la siguiente sección aplicaré estas y otras métricas secundarias[14] de la complejidad sintáctica contenidas en el sistema CLAS [Pakhomov y col.,

14 Estas métricas o índices secundarios se basan en la frecuencia de aparición de determinadas categorías gramaticales: nombres, verbos, etc.

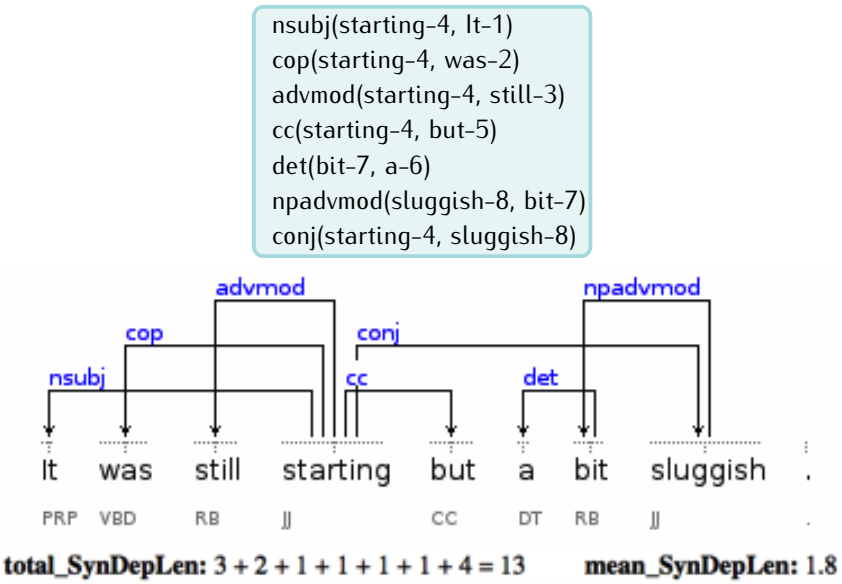


Figura 23: Cálculo del índice de Pakhomov.

2011] para clasificar las oraciones simétricas y asimétricas. Posteriormente, analizaré el rendimiento de un conjunto importante de clasificadores aplicados al cálculo de la polaridad pero omitiendo las oraciones asimétricas de los comentarios.

5.5.2 Clasificación automática de las oraciones simétricas y asimétricas

Este experimento busca determinar la validez de la **hipótesis 1**, formulada al principio de esta sección. Según esta hipótesis, las oraciones simétricas son sintácticamente menos complejas que las oraciones asimétricas. Por lo tanto, la hipótesis 1 puede reformularse como una pregunta: ¿es posible predecir satisfactoriamente cuándo una oración descriptiva es simétrica o asimétrica?

Para responder a esta pregunta analizo la complejidad sintáctica de las oraciones en el corpus MDTOD [Cruz Mata, 2012] empleando el sistema CLAS (*Computerized Linguistic Analysis System*) desarrollado por Pakhomov y col. [2011] (ver secciones 3.2.2 y 3.3.1.4). Como se puede observar en el Cuadro 48, el corpus MDTOD permite recuperar fácilmente las oraciones simétricas de las asimétricas contrastando la polaridad general del comentario (*rating*)[15] con la polaridad de cada una de sus oraciones (*opinión polarity*). Para no introducir ruido en los datos y evitar las inconsistencias, se han omitido del análisis las oraciones del corpus MDTOD que no expresan polaridad, las anotadas con doble polaridad (ej. *It is a [nice hotel]*$_{\texttt{positivo}}$, but [very expensive]$_{\texttt{negativo}}$), las que han sido segmentadas de forma errónea[16] y las que están formadas por una única palabra («ok», «duuuhhh»). El número total de oraciones seleccionadas por dominio aparece en el Cuadro 49.

15 Un *rating* igual o inferior a 2 representa un comentario negativo y uno igual o superior al 4, un comentario positivo. Los comentarios con una «polaridad» neutra (*rating* = 3) no se incluyeron en el análisis.

16 Por ejemplo, «oraciones» excesivamente extensas y que contenían determinados signos de puntuación («;», «.», «(», etc.).

```
<?xml version="1.0" encoding="UTF-8"?>
<review id="551" item="Bardstown Days Inn" rating="2">
  ...
 <text>
  <sentence id="1">
   Nice(1) enough(2) staff(3) at(4) this(5) hotel(6) right(7) near(8)
       the(9) Parkway(10) ...(11) but(12) the(13) upkeep(14) of(15)
       the(16) property(17) has(18) been(19) to(20) say(21) the(22)
       least(23) ...(24) lax(25) .(26)
   asimétrica <opinion polarity="+" feature="staff" featWords="3"
       opWords="1,2"/>
  </sentence>
  <sentence id="2">
   Water(1) damage(2) in(3) several(4) of(5) the(6) rooms(7) .(8)
  </sentence>
  <sentence id="3">
   Floors(1) that(2) are(3) buckled(4) .(5)
   simétrica <opinion polarity="-" feature="grounds" featWords="1"
       opWords="4"/>
  </sentence>
  ...
 </text>
</review>
```

Cuadro 48: Fragmento de un comentario del corpus MDTOD con oraciones simétricas y asimétricas.

	Coches	Auriculares	Hoteles	MDTOD
Simétricas	403	194	334	931
Asimétricas	403	194	334	931
Total	806	388	668	1862

Cuadro 49: Número total de oraciones simétricas y asimétricas seleccionadas en cada dominio.

Sobre este conjunto de oraciones he aplicado el sistema CLAS. Las quince métricas que se obtienen por cada oración con este sistema, entre las que se encuentran los índices de Yngve, Frazier y Pakhomov, fueron empleadas como atributos para entrenar y evaluar diferentes algoritmos de clasificación en Weka [Witten y col., 1999]. El Cuadro 50 contiene la lista completa de índices primarios y secundarios que calcula el sistema CLAS. Los índices o atributos primarios (1 al 6) se basan en la suma y los promedios de los valores individuales asociados a la profundidad o distancia entre palabras. Los índices o atributos secundarios (7 al 15) recogen la frecuencia de aparición de las palabras según su categoría gramatical. El principio que subyace al empleo de estos índices secundarios es que las oraciones complejas sintácticamente son más largas que las simples.

Con el objetivo de determinar la fiabilidad del sistema CLAS, calculé manualmente los 15 índices en 30 oraciones (10 por cada dominio) seleccionadas de manera aleatoria. Para determinar el nivel de acuerdo entre los valores obtenidos de forma automática y manual, apliqué el coeficiente Kappa [Cohen, 1960]. El resultado de esta prueba dio un valor `Kappa` más que aceptable, de 0.76, lo que valida el uso de esta herramienta para el análisis. Con el fin de normalizar los datos, realicé una transformación lineal de los índices originales obtenidos con CLAS para situarlos en una mismo rango de valores `[0..1]`. El paquete `ppls` [Krämer y Sugiyama, 2011] del entorno R fue usado con este propósito. Por su parte, la clasificación se llevó a cabo mediante validación cruzada de 10 iteraciones (10-fold cross-validation) que es la técnica

N	Índices / Atributos
1	Media de la «profundidad» de las palabras según el índice de Frazer (ver *mean_Fdepth* en la Figura 22).
2	Suma de la «profundidad» de las palabras según el índice de Frazer (ver *total_Fdepth* en la Figura 22).
3	Media de la «profundidad» de las palabras según el índice de Yngve (ver *mean_Ydepth* en la Figura 21).
4	Suma de la «profundidad» de las palabras según el índice de Yngve (ver *total_Ydepth* en la Figura 21).
5	Media de la distancia entre palabras con dependencia sintáctica según el índice de Pakhomov (ver *mean_SynDepLen* en la Figura 23).
6	Suma de la distancia entre palabras con dependencia sintáctica según el índice de Pakhomov (ver *total_SynDepLen* en la Figura 23).
7	Número de nodos «S» en el árbol de dependencias sintácticas.
8	Número de sustantivos.
9	Número de adjetivos.
10	Número de adverbios.
11	Número de verbos.
12	Número de determinantes.
13	Número de conjunciones.
14	Número de preposiciones.
15	Número de nombres propios.

Cuadro 50: Lista de los quince índices / atributos generados por el sistema CLAS (*Computerized Linguistic Analysis System*).

más extendida para evaluar los resultados de un análisis y garantizar que son independientes de la partición entre datos de entrenamiento y prueba.

El Cuadro 51 presenta los niveles de precisión conseguidos tras la clasificación de las oraciones del corpus MDTOD. La precisión indica el porcentaje de oraciones correctamente clasificadas entre todas las oraciones a las que se les asignó uno de los dos atributos de clase posibles: `simétrica` o `asimétrica`. En la parte central del cuadro están las precisiones conseguidas con los once clasificadores que fueron entrenados para cada dominio, por separado y en conjunto (es decir, el corpus MDTOD en su totalidad). En la parte baja del cuadro aparece el promedio de todas las precisiones.

Según el Cuadro 51, la mayoría de las medidas de precisión superan el 70 % de acierto para esta tarea de clasificación. Por lo tanto, los resultados de este experimento indican que la complejidad sintáctica es útil para predecir si una oración es simétrica o asimétrica, es decir, si tiene la misma polaridad o una polaridad opuesta a la del comentario. El mejor clasificador en el dominio de los coches es `DTNB` (`Decision Table Naive Bayes`) con un 72.3 % de acierto en la tarea propuesta. `DTNB` [Hall y Frank, 2008] es un clasificador híbrido que utiliza tanto tablas de decisión como redes bayesianas: en cada punto de la búsqueda, el algoritmo divide los atributos en dos conjuntos disjuntos, uno para la tabla de decisión y el otro para la red bayesiana. El mejor clasificador en el dominio de los auriculares es `ADTree` con un 73.7 % de acierto. `ADTree` [Freund y L. Mason, 1999] es una generalización de los árboles de decisión que combina métodos de decisión y predicción (*boosting*). El mejor clasificador en el dominio de los hoteles es `SimpleCart` con un 72.2 % de acierto. `SimpleCart` [Breiman y col., 1984] es una técnica de aprendizaje basada en árboles de decisión no paramétricos que proporcio-

	Coches	Auriculares	Hoteles	MDTOD
BayesNet	70.3	70.1	67.2	73.1
LWL	65.0	71.9	66.5	76.3
DTNB	**72.3**	70.6	69.5	75.8
Decis.Table	71.5	72.7	71.0	76.0
JRip	70.3	70.6	68.7	76.3
Ridor	71.7	70.4	69.9	76.4
ADTree	70.6	**73.7**	71.1	76.0
BFTree	71.7	71.1	70.4	76.4
LADTree	71.3	72.9	71.7	75.5
REPTree	70.7	71.6	69.5	75.5
SimpleCart	70.7	71.6	**72.2**	76.3
Promedio	*70.6*	*71.6*	*69.8*	*75.8*

Cuadro 51: Porcentaje de oraciones simétricas y asimétricas correctamente clasificadas en cada dominio.

na árboles de regresión o clasificación según de si la variable dependiente[17] es categórica o numérica, respectivamente. En general, los comentarios sobre hoteles tienden a marcar menos la diferencia entre oraciones simétricas y asimétricas (69.8 % de precisión promedio); esta diferencia es más marcada en los comentarios sobre auriculares (71.6 % de precisión promedio). Por último, la precisión más alta se consigue unificando todos los dominios del corpus MDTOD: en la última columna del Cuadro 51, todos los valores superan el 73 % de acierto.

Adicionalmente, con el objetivo de determinar cuáles son los atributos más relevantes[18] para diferenciar ambas clases de oraciones, he aplicado cuatro métodos estándar de ponderación de atributos que proporcionan una lista de atributos ordenada según su relevancia[19]:

- Chi-cuadrado (`Chi-squared`, χ^2): fue presentado como método de discretización pero más tarde se mostró que era capaz de eliminar atributos redundantes o no relevantes.
- Razón de ganancia (`Gain Ratio`, GR): es una medida que tiene en cuenta tanto la ganancia de información como las probabilidades de los distintos valores del atributo.
- Ganancia de información (`Information Gain`, IG): no es más que la reducción esperada de la entropía (desorden) de los datos al conocer el valor de un atributo.
- `ReliefF` (RL): es un algoritmo inspirado en el aprendizaje basado en casos que intenta obtener los atributos estadísticamente más relevantes hallando los vecinos más cercanos a las instancias que analizan.

17 La variable dependiente es la variable que se intenta predecir.

18 Aunque en Aprendizaje Automático no existe acuerdo sobre lo que debe entenderse por «relevante» ya que un atributo siempre es relevante *respecto al objetivo* que se asume de clasificación, una definición formal que se adapta al esquema de selección de atributos como un proceso de búsqueda es la siguiente: «Dado un conjunto de aprendizaje $\mathcal{E}$, un algoritmo de aprendizaje $\mathcal{L}$, y un conjunto de atributos $\mathcal{S}$, un atributo X_i es incrementalmente útil para $\mathcal{L}$ con respecto a $\mathcal{S}$ si la tasa de acierto de la hipótesis que $\mathcal{L}$ produce usando el conjunto de atributos $\{X_i\} \cup \mathcal{L}$ es mejor que la tasa de acierto obtenida utilizando sólo el conjunto de atributos $\mathcal{S}$» [Ruiz Sánchez, 2005].

19 Los evaluadores de atributos deben aplicarse con un método de búsqueda como `Ranker` o `BestFirst`.

En el Cuadro 52 aparecen los cinco atributos que predicen mejor cuándo una oración es simétrica o asimétrica en cada uno de los tres dominios analizados. En general, los índices que miden la profundidad o la distancia entre palabras (ej. `total_Ydepth`, mean_Ydepth, etc.) son más informativos que los índices basados en la frecuencia de aparición de las palabras (ej. `num_clauses`, `noun_count`, etc.). En particular, los índices que suman la profundidad o distancia de las palabras (`total_Ydepth`, `total_Fdepth` y `total_SynDepLen`) funcionan mejor en todos los dominios.

Coches			
χ^2	**GR**	**IG**	**RL**
total_SynDepLen	total_Ydepth	total_SynDepLen	total_SynDepLen
total_Ydepth	total_SynDepLen	total_Ydepth	total_Ydepth
total_Fdepth	det_count	total_Fdepth	total_Fdepth
mean_Ydepth	total_Fdepth	mean_Ydepth	mean_Ydepth
mean_SynDepLen	mean_Ydepth	mean_SynDepLen	mean_SynDepLen
Auriculares			
χ^2	**GR**	**IG**	**RL**
total_Ydepth	total_Ydepth	total_Ydepth	total_Fdepth
total_SynDepLen	total_SynDepLen	total_SynDepLen	total_Ydepth
total_Fdepth	mean_Ydepth	total_Fdepth	total_SynDepLen
mean_Ydepth	total_Fdepth	mean_Ydepth	adj_count
num_clauses	num_clauses	num_clauses	conj_count
Hoteles			
χ^2	**GR**	**IG**	**RL**
total_Ydepth	total_SynDepLen	total_Ydepth	total_Fdepth
total_SynDepLen	total_Ydepth	total_SynDepLen	num_clauses
noun_count	mean_Ydepth	mean_Ydepth	verb_count
mean_Ydepth	noun_count	noun_count	mean_Ydepth
total_Fdepth	total_Fdepth	total_Fdepth	total_Ydepth

Cuadro 52: Atributos más relevantes recuperados por los métodos de selección de atributos en cada uno de los dominios analizados.

5.5.3 Cálculo de la polaridad mediante el segmento descriptivo

Al principio de esta sección he asumido que el segmento descriptivo está constituido por el conjunto de las oraciones que en un comentario «describen» positiva o negativamente un producto o alguna de sus características. Estas oraciones las he clasificado como simétricas o asimétricas según expresan la misma polaridad o la polaridad opuesta a la del comentario completo. El experimento que presento a continuación busca determinar la validez de la **hipótesis 2**, formulada en la Sección 5.5.1. Según esta hipótesis, es posible mejorar los niveles de precisión en el cálculo de la polaridad omitiendo las oraciones que expresan una polaridad opuesta a la del comentario completo (oraciones asimétricas). Por lo tanto, la hipótesis 2 puede reformularse como una pregunta: ¿es posible mejorar el rendimiento de los algoritmos de clasificación de la polaridad eliminando las oraciones asimétricas de los comentarios sobre productos? Para responder a esta pregunta he calculado la polaridad de los comentarios mediante SO-CAL [Taboada, Brooke, Tofiloski y col., 2011] contrastando tres configuraciones diferentes de los datos de entrada. Todas las configuraciones se basan en la omisión de algún fragmento del comentario. Las configuraciones propuestas son:

BASELINE: En esta configuración el análisis de la polaridad se llevó a cabo de manera convencional, es decir, usando el comentario completo. La entrada al SO-CAL está constituida por todas las oraciones del texto, tanto las que expresan polaridad (positiva o negativa) como las que no expresan ninguna polaridad. A continuación, un ejemplo del tipo de entrada que recibe el SO-CAL bajo esta configuración (el comentario es negativo):

(17)  Our first stay in NY one week in january. This hotel is located well in Soho. Perfect located in the Village, central but a more quiet place. Superfriendly staff. Service was on point. The lobby smells wonderful, like a spa. The rooms are spacious enough. The bathrooms were super luxurious. Nice amenities. The breakfast was nice. The breakfast room small and cosy. The only problem was, however, that while we were in Central Park, our room was robbed!! The thieves stole 2 laptops and cash. Many cases like ours have been reported on TripAdvisor. I probably will not return though.

GOLD STANDARD: En esta configuración el análisis de la polaridad se llevó a cabo omitiendo todas las oraciones asimétricas de los comentarios. La entrada al SO-CAL está constituida por las oraciones que expresan la misma polaridad que la del comentario conjuntamente con las oraciones que no expresan ninguna polaridad. Esta configuración corresponde al *gold standard* puesto que la selección del tipo de oración se basa en una predicción hipotética e ideal del 100 % de acierto en la clasificación de oraciones simétricas y asimétricas. A continuación, un ejemplo de este tipo de entrada (el fragmento del comentario que ha sido omitido aparece tachado):

(18)  Our first stay in NY one week in january. ~~This hotel is located well in Soho. Perfect located in the Village, central but a more quiet place. Superfriendly staff. Service was on point. The lobby smells wonderful, like a spa. The rooms are spacious enough. The bathrooms were super luxurious. Nice amenities. The breakfast was nice. The breakfast room small and cosy.~~ The only problem was, however, that while we were in Central Park, our room was robbed!! The thieves stole 2 laptops and cash. Many cases like ours have been reported on TripAdvisor. I probably will not return though.

PROPUESTA: En esta configuración el análisis de la polaridad se llevó a cabo omitiendo algunas de las oraciones asimétricas de los comentarios. El porcentaje de oraciones asimétricas omitidas se realizó de acuerdo con los resultados obtenidos en el experimento anterior (ver promedios en el Cuadro 51). La entrada al SO-CAL está constituida por todas las oraciones del comentario excepto el 70 % de oraciones asimétricas en el dominio de los coches, el 71 % de oraciones asimétricas en el dominio de auriculares y el 69 % de oraciones asimétricas en el dominio de hoteles. Se elimina estos porcentajes porque, como hemos visto, son los que es capaz de detectar automáticamente el clasificador. En todos los casos la selección de las oraciones asimétricas se efectuó de forma aleatoria. A continuación, un ejemplo de este tipo de entrada (el fragmento del comentario que ha sido omitido aparece tachado):

(19)  Our first stay in NY one week in january. ~~This hotel is located well in Soho.~~ Perfect located in the Village, central but a more quiet place. ~~Superfriendly staff.~~ ~~Service was on point.~~ The lobby smells wonderful, like a spa. ~~The rooms are spacious enough.~~ ~~The bathrooms were super luxurious.~~ ~~Nice amenities.~~ The breakfast was nice. ~~The breakfast room small and cosy.~~ The only problem was, however, that while we were in Central Park, our room was robbed!! The thieves stole 2 laptops and cash. Many cases like ours have been reported on TripAdvisor. I probably will not return though.

Los resultados de este experimento se presentan en el Cuadro 53. En la segunda columna aparece la configuración a superar, es decir, el *baseline*. Los niveles de precisión obtenidos mediante esta configuración de base son consistentes con los obtenidos por otros investigadores para la misma tarea (ver Graovac y Pavlovic-Lazetic [2014]). En la tercera columna se exhiben los valores de referencia, es decir, el *gold standard* (valores ideales). En la cuarta columna (*Propuesta*) presento las precisiones obtenidas mediante mi aproximación.

Configuración	Baseline		Gold standard		Propuesta	
Comentario (pol.)	+	−	+	−	+	−
Coches	88.1	77.3	93.4	85	89.4	83.2
Auriculares	81	74.3	89.2	80.6	83.3	78.7
Hoteles	95	78.8	97	88	96.7	81.9
Promedio	88	76.8	93.2	84.5	89.8	81.2

Cuadro 53: Análisis de la polaridad de los comentarios sobre Coches, Hoteles y Auriculares (mediante SO-CAL) bajo diferentes configuraciones.

Como se puede observar, las precisiones obtenidas eliminando las oraciones asimétricas que los clasificadores aciertan a identificar basándose en su complejidad sintáctica (columna `Propuesta`), están más próximas al *gold standard* que las del *baseline*. Por ejemplo, mientras que el *gold standard* predice los comentarios positivos en el dominio de los Auriculares con un 89.2 % de acierto, el *baseline* los predice en un 81 %, es decir, 8.2 puntos de diferencia en relación al *gold standard*. No obstante, la precisión obtenida para el mismo tipo de comentarios y en el mismo dominio según la *propuesta* es de 83.3 %, es decir, 5.9 puntos de diferencia en relación al *gold standard* y 2.3 puntos por encima del *baseline*.

Esta mejora en el rendimiento de los clasificadores producida por la omisión de las oraciones asimétricas es más notoria en el caso de los comentarios negativos. En efecto, la omisión de las oraciones con una polaridad opuesta a la del comentario tiene mayores repercusiones sobre los comentarios con polaridad negativa. Por ejemplo, mientras que el *gold standard* predice los comentarios negativos en el dominio de los coches con un 85 % de acierto, el *baseline* los predice en un 77.3 %, es decir, 7.7 puntos de diferencia en relación al *gold standard*. Por su parte, la precisión obtenida según la *Propuesta* es de 83.2 %, es decir que está tan sólo a un 1.8 puntos de diferencia en relación al *gold standard*, 5.9 puntos mejor que el *baseline*. Los promedios generales de precisión también constatan esta divergencia entre comentarios positivos y negativos. Si contrastamos el *baseline* y la *Propuesta* tenemos que la precisión pasa del 88 % al 89.8 % en el caso de los comentarios positivos y del 76.8 %

al 81.2 % en el de los comentarios negativos. En el primer caso se mejora el rendimiento en un 1.8 y en el segundo en un 4.4.

Por lo tanto, los resultados indican que diferenciar entre oraciones descriptivas simétricas y oraciones descriptivas asimétricas es útil para detectar la polaridad de los comentarios: al omitir las oraciones sintácticamente complejas (oraciones asimétricas) de estos textos de opinión mejoramos el rendimiento de los clasificadores que predicen su polaridad. Adicionalmente, y esto es lo más relevante del análisis, se ha detectado una divergencia importante entre el tipo de polaridad y el uso del segmento descriptivo.

5.6 FUNCIÓN DE LOS SEGMENTOS DISCURSIVOS EN LA EXPRESIÓN DE LA POLARIDAD

En las secciones anteriores he presentado los resultados de varios experimentos mediante los cuales identifico los tres tipos de segmentos que componen los comentarios sobre productos y los empleo para analizar la polaridad del comentario. Gracias a haber contrastado el efecto que produce el uso o la omisión de un determinado segmento sobre el cálculo de la polaridad, ha sido posible determinar la función que cada segmento desempeña en la expresión de la polaridad. En este apartado establezco dicha función a la luz de los experimentos presentados a lo largo del capítulo.

El Cuadro 54 resume, en términos cuantitativos, la relación que se ha establecido entre la polaridad de los comentarios sobre productos y los diferentes tipos de segmentos discursivos que los conforman. En el cuadro se compara la precisión obtenida al calcular la polaridad mediante «todo el comentario» con la obtenida a partir del uso o la omisión de un determinado segmento. En la comparativa también se tienen en consideración la aproximación empleada, el idioma de los comentarios y su polaridad. En la parte izquierda del cuadro aparece el nombre del segmento que ha podido ser identificado y aplicado satisfactoriamente al análisis de la polaridad mediante el uso de diferente información lingüística: información morfosintáctica y léxica, secuencias narrativas y complejidad sintáctica.

La parte superior del Cuadro 54 corresponde a los experimentos presentados en la Sección 5.3 mediante los cuales ha sido posible identificar la función que ddesempeña el segmento valorativo en la expresión de la polaridad de los comentarios. En estos experimentos se emplea información morfosintáctica y léxica para segmentar los comentarios en español e inglés, y dos aproximaciones para calcular la polaridad: BoW y SO-CAL. La primera aproximación se basa en aprendizaje supervisado puesto que se aplican una serie de algoritmos que tienen en consideración la polaridad de los comentarios en el momento de aprender a clasificarlos. Bajo BoW, la precisión obtenida a partir del segmento valorativo es similar o superior a la conseguida empleando todo el comentario, tanto en español como en inglés. Estos resultados me permiten afirmar que los segmentos valorativos expresan la polaridad de los comentarios de manera más efectiva que el comentario entero o que los otros segmentos de forma aislada usando BoW. El buen desempeño de BoW-S_v permite afirmar que el segmento valorativo tiene la función de sintetizar o expresar la polaridad general del comentario. Por lo tanto, si aislamos el segmento valorativo, podemos obtener la polaridad del cometario con un nivel de precisión superior o, por lo menos, similar a la precisión obtenida

VALORATIVO

INFORMACIÓN MORFOSINTÁCTICA Y LÉXICA				
Aproximación	Idioma	todo el comentario	usando el segmento valorativo	pol
BoW	Esp	71.6 %	70 %*	+/-
(supervisado)	Ing	78.4 %	79.3 %**	+/-
SO-CAL (no supervisado)	Ing	87.9 %	72.2 %**	+/-

Estos valores proceden de los Cuadros 39 y 40.
*Se considera una mejora en el rendimiento puesto que esta presición fue obtenida usando tan sólo un 10.8 % del texto del comentario.
**Esta presición fue obtenida usando un 18 % del texto del comentario.

NARRATIVO

SECUENCIAS NARRATIVAS				
Aproximación	Idioma	todo el comentario	sin secuencias narrativas	pol
SO-CAL	Ing	89.3 %	87.3 %	+
(no supervisado)	Ing	75.7 %	65.4 %	-

Estos valores corresponden a la precisión promedio de las representaciones C_{ref} y des_{ext} para los comentarios positivos (+) y negativos (–) descritos en el Cuadro 47.

DESCRIPTIVO

COMPLEJIDAD SINTÁCTICA				
Aproximación	Idioma	todo el comentario	sin descripciones asimétricas	pol
SO-CAL	Ing	88 %	89.8 % (93.2 %*)	+
(no supervisado)	Ing	76.8 %	81.2 % (84.5 %*)	-

Estos valores proceden del Cuadro 53.
*Corresponden al *gold standard*.

Cuadro 54: Efectos del uso de los segmentos discursivos sobre el cálculo de la polaridad de los comentarios.

usando el comentario completo. Esta relación entre el segmento valorativo y la expresión de la polaridad queda evidenciada en el comentario (20).

(20) ★ ★ ★ ★ ★

Sabiamos que el hotel estaba bien porque lo vimos en Internet pero cuando llegamos nos sorprendio todo exquisitamente limpio, todo tipo de detalles, por ponerle un pero la ubicacion no viene muy bien explicada. Simplemente nos quedamos con la boca abierta. De corazon os lo recomiendo, volveria a ir.

La segunda aproximación se basa en aprendizaje no supervisado puesto que SO-CAL solo tiene en consideración la polaridad de base de las palabras a la hora de clasificar los comentarios. Los experimentos mostraron una clara diferencia en el rendimiento de los clasificadores determinada por la **aproximación**: el rendimiento decae notablemente cuando la polaridad del comentario se calcula mediante la herramienta SO-CAL. Este hecho me ha llevado a plantear procedimientos alternativos para el tratamiento de los segmentos narrativo y descriptivo.

La parte media del Cuadro 54 corresponde a los experimentos presentados en la Sección 5.4 mediante los cuales ha sido posible identificar la función que ddesempeña el segmento narrativo en la expresión de la polaridad de

los comentarios. En estos experimentos se han identificado las secuencias narrativas de los comentarios como paso previo al análisis de la polaridad. Al seleccionar el segmento narrativo de los comentarios ha resultado útil para determinar su función. En primer lugar, al omitir las secuencias narrativas se obtuvo una mejora significativa sobre la precisión alcanzada mediante el segmento valorativo bajo SO-CAL: SO-CAL-S_v = 72.2 % y SO-CAL-S_n = 76.3 %[20]. En segundo lugar, los resultados evidenciaron que la omisión de los segmentos narrativos –representados, en este caso, por las secuencias narrativas– afectaba casi de manera exclusiva a los comentarios con polaridad negativa: al omitir el segmento narrativo en los comentarios con polaridad positiva se ha pasado de un 89.3 % a un 87.3 % de precisión (2.0 puntos de diferencia) mientras que al omitir el mismo segmento en los comentarios con polaridad negativa se ha pasado de una precisión del 75.7 % a una del 65.4 % (10.3 puntos de diferencia). Este resultado me permite afirmar que los usuarios recurren a las narraciones para comentar aspectos negativos del producto valorado como un mecanismo transversal de expresión de la polaridad. Por lo tanto, los segmentos narrativos además de contextualizar la opinión en los comentarios, tienen la función específica de expresar la polaridad negativa. En otras palabras, la narración se usa como mecanismo lingüístico para expresar una opinión negativa. El comentario (21) sirve para ejemplificar esta relación entre segmentos narrativos y polaridad negativa:

(21) ★ ★ ★ ★ ★

El c4 desde mi punto de vista no es un coche muy bueno para mi opinión simplemente, por q un amigo mio se dio una hostia bastante pekeña y partio la dirección (eso un coche duro no le pasaría asi por ke si, con una minima hostia). Pero si nos ponemos a hablar de comodidad y de mas desde mi punto de vista, el coche es cómodo y seguro (en algunos aspectos, por ejemplo airbag y de mas son buenos y seguros)...

Finalmente, la parte baja del Cuadro 54 corresponde a los experimentos presentados en la Sección 5.5 mediante los cuales ha sido posible identificar la función que desempeña el segmento descriptivo en la expresión de la polaridad de los comentarios. En estos experimentos se han seleccionado las oraciones descriptivas simétricas y asimétricas de los comentarios para efectuar al análisis de la polaridad. En general, diferenciar entre oraciones simétricas y asimétricas ha resultado útil para detectar la polaridad de los comentarios. En primer lugar, al omitir las oraciones asimétricas –caracterizadas por ser sintácticamente complejas– se ha conseguido superar la precisión de cualquiera de las anteriores representaciones: 89.8 % para los comentarios positivos y 81.2 % para los negativos (85.5 % en promedio). Desde el punto de vista lingüístico, este hallazgo nos está indicando que los usuarios suelen hacer uso de estructuras sintácticamente complejas para expresar las opiniones individuales con una polaridad opuesta a la del comentario. Este fenómeno se acentúa en el caso de los comentarios con polaridad negativa puesto que, por una cuestión de cortesía lingüística[21], resulta más viable para el usuario evaluar negativamente un producto en términos positivos que negativos. El comentario (22) sirve para ejemplificar esta relación entre las oraciones asimétricas o sintácticamente complejas y la polaridad del comentario (en este caso positiva):

20 SO-CAL-S_n = (87.3 % + 65.4 %) / 2 = 76.3 %.

21 El fenómeno de la cortesía lingüística y su relación con la expresión de la polaridad en comentarios ha sido tratado por Ortiz y col. [2010].

(22) ★ ★ ★ ★ ★

Un hostal muy limpio, comodo y bien ubicado. La cocina es amplia y muy bien equipada. Los espacios comunes muy lindos y confortables. Si tengo que comentar algo negativo, me inclino por las habitaciones, que son algo pequeñas y con pocas medidas de seguridad...

En general, los experimentos referidos en este capítulo han revelado que, a excepción del segmento valorativo, los segmentos narrativo y descriptivo exhiben funciones diferentes o modifican su estructura según se trate de un comentario positivo o negativo. El segmento narrativo, que en los comentarios con polaridad positiva tiene la función (manifiesta) de contextualizar la opinión, en los comentarios con polaridad negativa se usa para expresar dicha polaridad. El segmento descriptivo, que en los comentarios con polaridad positiva suele describir los productos mediante estructuras simples, modifica su comportamiento para exhibir una estructura compleja en los comentarios con polaridad negativa.

5.7 CONCLUSIONES

En este capítulo he analizado la función que desempeñan los segmentos valorativo, narrativo y descriptivo en el cálculo de la polaridad de los comentarios sobre productos.

La Sección 5.2 está dedicada a la clasificación automática de los tres tipos de segmentos que conforman los comentarios sobre productos. Los experimentos que llevé a cabo en esta primera parte del análisis me permitieron determinar que es posible aislar de forma automática, con una precisión promedio del 80 %, los segmentos valorativo, narrativo y descriptivo mediante el uso de un conjunto de características léxicas y morfosintácticas. Adicionalmente, he podido comprobar que los segmentos valorativos expresan la polaridad de los comentarios de manera más efectiva que el comentario entero o que los otros segmentos de forma aislada bajo una aproximación supervisada (BoW-S_v).

La Sección 5.4 está dedicada al análisis del segmento de tipo narrativo. Los experimentos que llevé a cabo en esta segunda parte del análisis de la polaridad me permitieron determinar que es posible recuperar de forma automática el segmento narrativo seleccionando las secuencias narrativas que forman parte de los comentarios. Además, he observado que los usuarios recurren a las narraciones para comentar aspectos negativos del producto valorado como un mecanismo transversal de expresión de la polaridad.

La Sección 5.5 está dedicada al análisis del segmento de tipo descriptivo. Los experimentos que llevé a cabo en esta última parte del análisis de la polaridad me permitieron constatar que es posible usar la complejidad sintáctica para diferenciar las oraciones que expresan la misma polaridad o la polaridad opuesta a la del comentario. También he observado que la omisión de las oraciones asimétricas (representadas por las oraciones sintácticamente complejas) mejora la detección de la polaridad de los comentarios, especialmente la de los comentarios con polaridad negativa. Este hecho indica que los usuarios se suelen valer de las estructuras sintácticamente complejas para expresar opiniones con una polaridad opuesta a la del comentario.

6 CONCLUSIONES, CONTRIBUCIONES Y TRABAJO FUTURO

En este capítulo recojo los principales resultados obtenidos en la investigación. La presentación de tales resultados la realizo en tres apartados diferentes: conclusiones (Sección 6.1), contribuciones (Sección 6.2) y trabajo futuro (Sección 6.3).

6.1 CONCLUSIONES

En esta tesis he tratado el problema de la detección de la polaridad en comentarios sobre productos con el fin de mejorar los modelos de recomendación. Para predecir la polaridad de este tipo de textos de opinión he propuesto un enfoque basado en el análisis de su estructura discursiva. La primera parte de mi trabajo comienza con la caracterización de los comentarios como un género discursivo (Capítulo 4). A partir de dicha caracterización ha sido posible identificar los principales tipos de segmentos que forman parte del género discursivo de los comentarios sobre productos. La segunda parte de mi trabajo ha consistido en identificar y utilizar cada tipo de segmento para analizar la polaridad de los comentarios (Capítulo 5). El análisis lingüístico de cada uno de los tipos de segmentos me ha servido para identificar la función que desempeñan en la expresión de la polaridad.

6.1.1 Caracterización del género discursivo

El análisis y los experimentos asociados a la caracterización del género discursivo revelaron que los comentarios sobre productos se componen de tres tipos básicos de segmentos: narrativo, descriptivo y valorativo. Atendiendo a su propósito comunicativo, la función del segmento narrativo es contextualizar la opinión mediante la mención de los eventos asociados a la valoración de un producto; la función del segmento descriptivo es caracterizar el producto mediante la mención de sus rasgos o propiedades más relevantes; y la función del segmento valorativo es recomendar o no el producto mediante la expresión de la actitud personal del usuario respecto del producto valorado.

Adicionalmente, los experimentos realizados han revelado que, además de una estructura discursiva estable, los comentarios también comparten un mismo registro lingüístico. El análisis lingüístico determinó que si se comparan los comentarios con textos con un registro claramente formal como son los artículos periodísticos, las diferencias léxicas y morfosintácticas entre ambas clases de textos son significativas. Por el contrario, el mismo análisis lingüístico aplicado a los comentarios agrupados según criterios objetivos-demográficos (sexo, edad y procedencia de los usuarios) no permitió identificar diferencias remarcables. Estos hechos indican que los comentarios forman parte de un conjunto indisociable de textos, el del género discursivo de los comentarios sobre productos.

6.1.2 Análisis de la polaridad

El análisis y los experimentos con el análisis de la polaridad han revelado que es posible clasificar de forma automática los diferentes tipos de segmentos discursivos que conforman los comentarios. Estos experimentos también revelan que cada tipo de segmento contribuye de forma diferente a la expresión de la polaridad. A continuación sintetizo el procedimiento que he aplicado para analizar la polaridad de los comentarios.

1. En primer lugar, mediante el uso de un conjunto restringido de rasgos lingüísticos de fácil obtención he podido identificar, de forma automática, los tres tipos de segmentos que forman parte del género discursivo de los comentarios sobre productos: segmento narrativo, descriptivo y valorativo. En estos experimentos usé, con muy buenos resultados, 14 rasgos obtenidos etiquetando con información morfosintáctica los textos y generando o aplicando varios léxicos.

2. En segundo lugar, he empleando los segmentos discursivos para calcular la polaridad del comentario aplicando tanto aprendizaje supervisado (BoW) como no supervisado (SO-CAL), con diferentes resultados. Al aplicar aprendizaje supervisado, basado en bolsa de palabras, he comprobado que el segmento valorativo presenta un rendimiento similar al que se obtiene usando todo el texto del comentario pero con una reducción de los datos de entrada cercana al 90 %. Estos resultados me han llevado a concluir que los usuarios emplean el segmento valorativo para expresar la polaridad general del comentario. Este segmento suele formularse en primera persona y ubicarse al final del comentario. Las implicaciones que estos hallazgos tienen para el análisis de la polaridad se pueden resumir en la necesidad de dar mayor prioridad y protagonismo a esta parte del comentario y en considerar su propósito comunicativo a la hora de establecer cualquier aproximación para la detección de la polaridad de los comentarios sobre productos. No obstante, al aplicar aprendizaje no supervisado mediante el uso de la herramienta SO-CAL, el rendimiento del segmento valorativo se reduce notablemente. Este hecho ha motivado la necesidad de entender el papel que cumplen los segmentos narrativo y descriptivo en la expresión de la polaridad. Las siguientes dos etapas de mi propuesta han estado encaminadas a analizar ambos segmentos.

3. En tercer lugar, mediante la selección de las oraciones cuyos verbos comparten argumentos correferentes he podido identificar las secuencias narrativas de los comentarios. Al aplicar dichas secuencias para calcular la polaridad del comentario constaté que, conjuntamente con la descripción de características negativas, los usuarios recurrían a las narraciones (relación de eventos) para manifestar su no conformidad con un producto. Este hallazgo indica que el segmento narrativo, además de utilizarse para introducir y contextualizar las opiniones, se usa con el propósito de expresar la polaridad negativa. Mi recomendación, en este sentido, es integrar el análisis del segmento narrativo, en particular, y de la narración, en general, en los estudios sobre el análisis de la polaridad. La narración es una tipología textual potencialmente útil para comprender y explicar la expresión de las opiniones negativas en lenguaje natural.

4. En cuarto lugar, mediante la aplicación de varias métricas de la complejidad sintáctica he podido clasificar con buenos resultados las oraciones

(descriptivas) que expresan la misma polaridad o una polaridad opuesta a la del comentario. Adicionalmente, al basarme en esta clasificación para calcular la polaridad del comentario, he constatado que la omisión de las oraciones con polaridad opuesta a la del comentario facilita, especialmente, la detección de la polaridad en los comentarios negativos. Este hecho me ha permitido concluir que los usuarios se suelen valer de las estructuras sintácticamente complejas para expresar opiniones asimétricas, es decir, contrarias a la polaridad general del comentario. Los resultados expuestos señalan la necesidad de incorporar la complejidad sintáctica en las investigaciones sobre el análisis de la polaridad.

En general, los diferentes análisis estructural y lingüístico de los comentarios me han permitido establecer diferencias notables asociadas a la expresión de la polaridad positiva y negativa. Esto se debe a que la valoración positiva es un tipo de polaridad «no marcada», es decir, la polaridad que determina la configuración (por defecto) del género discursivo. Las opiniones positivas son directas, claras y, en este sentido, no suelen emplear recursos como la ironía o el sarcasmo. En general, son fáciles de procesar a nivel cognitivo y, en cierta medida, a nivel computacional. Las opiniones negativas, por el contrario, no se suelen expresar de forma directa ya que podrían resultar violentas y desagradables. Aun tratándose de textos coloquiales, los usuarios intentan mantener las formas y las buenas maneras en sus comentarios.

Este cambio (*switching*) de una polaridad positiva «no marcada» a una polaridad negativa «marcada» comporta cambios formales y funcionales en los segmentos de los comentarios. El primer cambio se produce en la función de los segmentos narrativos. Si en los comentarios positivos los segmentos narrativos se utilizan para contextualizar y relatar experiencias, en los comentarios negativos se emplean, además, para expresar la polaridad de manera indirecta. Las narraciones se suelen presentar mediante secuencias de eventos organizados cronológicamente, es decir, mediante secuencias narrativas. El hecho de que se puedan utilizar secuencias narrativas para expresar la polaridad negativa obedece, a mi modo de ver, a que este tipo de secuencias están gobernadas por una lógica interna que facilita al usuario entender el propósito de los enunciados: todos los hablantes de una lengua estamos familiarizados con la lógica de la narración. La narración permite acercar al usuario gradualmente a un «desenlace negativo» que no es necesario en el caso de las valoraciones positivas. Por lo tanto, los segmentos narrativos producen el efecto deseado por el «comentarista» (*reviewer*) con un mínimo esfuerzo de procesamiento por parte del lector del comentario. El segundo cambio que se deriva de la expresión de opiniones negativas afecta a los segmentos descriptivos. En este caso no se ve comprometida la función de este tipo de segmentos (presentar las características de un producto) sino la de su estructura lingüística. Debido al mayor esfuerzo cognitivo que demanda generar una opinión negativa, las descripciones se tornan sintácticamente complejas, en un nivel que contrasta con la simplicidad de las descripciones en los comentarios positivos.

Considero que las diferencias detectadas entre opiniones positivas y negativas son lo suficientemente importantes como para permitirme afirmar que estamos ante dos tipos de comentarios o subgéneros discursivos, en lugar de uno: el subgénero de los comentarios positivos y el subgénero de los comentarios negativos. Cada subgénero tiene un propósito comunicativo particular y opuesto al de su contraparte: **valorar un producto positivamente** y **valorar un producto negativamente**. En este sentido, tanto la narración de eventos como la complejidad sintáctica serían mecanismos lingüísticos que

evidenciarían o «marcarían» este cambio en el propósito comunicativo de los comentarios sobre productos. Ambos subgéneros compartirían, entre otras cosas, un claro propósito persuasivo que converge en el uso del segmento valorativo. Las investigaciones en el ámbito del análisis de la polaridad, especialmente aquellas que se basan en la estructura discursiva, deberían tener presente esta doble naturaleza de los comentarios representada por la presencia de los segmentos narrativo, descriptivo y valorativo.

6.2 CONTRIBUCIONES

Esta tesis contribuye a mejorar la detección de la polaridad de los comentarios sobre productos en diferentes idiomas (inglés y español) y dominios (hoteles, coches y auriculares); aporta nuevos datos y nuevo conocimiento para entender la forma en que los usuarios evalúan productos en lenguaje natural; y contribuye a proporcionar soluciones a otros problemas relacionados con el tratamiento de las opiniones como son la expresión de la polaridad negativa. De manera detallada, las principales contribuciones expuestas en la presente memoria son las siguientes:

1. La implementación de una herramienta (AToP) para clasificar los comentarios según diversos atributos demográficos mediante el cálculo de la riqueza léxica de los textos.

2. La compilación de un corpus de comentarios sobre hoteles (HOpinion) anotado con información morfosintáctica y a nivel de segmentos, que ha sido utilizado en otras investigaciones orientadas al tratamiento de la polaridad.

3. El reconocimiento de la necesidad de incorporar la estructura del discurso, a través del análisis del género discursivo, en el tratamiento de la polaridad de los comentarios sobre productos.

4. La caracterización de los comentarios como un género discursivo constituido por dos subgéneros: el subgénero de los comentarios positivos y el subgénero de los comentarios negativos.

5. La determinación de que las secuencias narrativas y la complejidad sintáctica están relacionadas con la expresión de la polaridad.

6. Un conjunto jerarquizado de rasgos léxicos y morfosintácticos que caracterizan el registro lingüístico de los comentarios sobre productos.

7. Una clasificación de las métricas de riqueza léxica más efectivas para determinar el registro lingüístico de los comentarios.

8. Una propuesta viable para segmentar los comentarios que resulta válida para diferentes dominios e idiomas puesto que se basa en los tipos textuales universales.

9. Una descripción de los principales tipos de segmentos que conforman la estructura de los comentarios: el orden canónico en el que suelen aparecer (etapas), combinatoria (*patterns*), opcionalidad y obligatoriedad de los mismos.

10. Un conjunto de rasgos lingüísticos que caracterizan cada uno de los tipos de segmentos discursivos que componen los comentarios sobre

productos y que permiten clasificar dichos segmentos de forma automática.

11. La identificación del propósito comunicativo asociado a los segmentos valorativo, narrativo y descriptivo.

12. El reconocimiento de que los segmentos discursivos cumplen propósitos comunicativos diferentes según la polaridad de los comentarios.

    Adicionalmente, como se detallan al principio de cada capítulo, parte de los logros obtenidos a lo largo de la tesis han sido publicados en revistas con proyección internacional.

6.3 TRABAJO FUTURO

Las líneas de trabajo futuro se dirigirán a:

1. Un primer punto de trabajo futuro es evaluar el rendimiento de la segmentación discursiva de los comentarios en otros idiomas y dominios. En este último caso, es importante tener presente que algunos dominios recurren más a la narración de eventos que otros (ej. comentarios sobre películas o libros) o pueden presentar una mayor complejidad estructural derivada del tratamiento de temas técnicos (ej. comentarios sobre el mercado de valores[1]).

2. Otro punto a tratar en un trabajo futuro tiene que ver con las métricas de la complejidad sintáctica. Dado que las métricas de la complejidad sintáctica existentes están pensadas para medir desordenes en el desarrollo del lenguaje o para determinar la competencia comunicativa de estudiantes de L2, considero indispensable desarrollar métricas de la complejidad sintáctica adaptadas al análisis de la polaridad en textos de opinión como los comentarios sobre productos. Estas métricas han de tener presente, entre otras cuestiones, que los comentarios son un tipo de texto coloquial y que por lo tanto suele presentar una segmentación oracional inadecuada (o inexistente), errores tipográficos y ortográficos frecuentes o pueden ser traducciones automáticas de comentarios originales.

3. Basándome en los resultados de esta investigación y teniendo en consideración la distinción planteada por algunos autores entre descripciones subjetivas y objetivas, considero que la clasificación de los segmentos descriptivos en subjetivos y objetivos puede aportar nuevos datos para mejorar el análisis de la polaridad de los comentarios.

4. Por último, un tema que se ha manifestado a través del trabajo con los corpus pero que no fue posible tratar en esta investigación, es la posible relación entre negación y polaridad. Este es, sin duda, un asunto que merece un capítulo independiente en los estudios sobre el análisis de la polaridad.

1 *Stock market reviews.*

Anexo

A PLATAFORMA ATOP

En este anexo describo las características y funcionalidades de AToP, un *software* creado para facilitar el análisis de los comentarios sobre productos. La programación de AToP ha sido realizada por Bakary Singateh Queral como trabajo de fin de grado en la Facultad de Ingeniería Informática de la Universidad de Barcelona [Singateh, 2013]. Este trabajo fue dirigido por Maria Salamó y asesorado por John Roberto lo asesoró. El trabajo incluye los resultados de la investigación publicada en el artículo:

- John Roberto, Maria Salamó y M. Antònia Martí [2012], «Análisis de la riqueza léxica en el contexto de la clasificación de atributos demográficos latentes», *Procesamiento de Lenguaje Natural*, 1, 48, pág. 97-104

A.0.1 Descripción

AToP es una Plataforma[1] en Java para el Análisis de Textos de Opinión en Lenguaje Natural que se ha implementado con el objetivo de facilitar el análisis automático de comentarios sobre productos. La herramienta AToP permite calcular varias métricas de la riqueza léxica con las cuales realizar *tests* de clasificación basados en diferentes algoritmos de aprendizaje automático. AToP integra los siguientes recursos que aplica en el procesamiento de los comentarios:

1. *Corpus HOpinion:* Es el corpus de comentarios sobre hoteles usado en esta tesis.
2. *MySQL[2]/PostGreSQL[3]:* Son dos sistemas de gestión de bases de datos libres que se utilizan para almacenar el corpus HOpinion.
3. *Weka[4]:* Es un conjunto de librerías JAVA para la extracción de conocimientos desde bases de datos.
4. *XmlStarlet[5]:* Es un conjunto de utilidades que operan bajo línea de comandos que se usa para transformar, consultar, validar y editar archivos XML.

Los requisitos de *software* y *hardware* que son necesarios para poner en funcionamiento la herramienta AToP son los siguientes:

1. El proyecto se ha realizado con el entorno de programación *NetBeans* y todas las pruebas se han realizado bajo el sistema operativo Ubuntu 12.04.
2. Se ha de instalar la XmlStarlet para ejecutar los comandos de *shell* mediante los que se obtienen de los archivos *xml* todos los valores necesarios para calcular las métricas de la riqueza léxica.

1 Una plataforma es un sistema que integra y ejecuta determinadas aplicaciones para ofrecer una funcionalidad definida.
2 `https://www.mysql.com/`
3 `http://www.postgresql.org.es/`
4 `http://www.cs.waikato.ac.nz/ml/weka/`
5 `http://xmlstar.sourceforge.net/`

3. Se requiere 1 GB mínimo de memoria RAM necesarios para almacenar todos los valores necesarios para efectuar las pruebas de análisis.
4. Se requiere un procesador del al menos 1 Ghz necesario para calcular todas las métricas y para realizar las pruebas de minería de datos.

A.0.2 Funcionalidades

AToP se ha implementado siguiendo el patrón de arquitectura de software MVC (Modelo Vista Controlador). A continuación se presenta y describe el Diagrama de Casos de Uso que según el patrón MVC resume el comportamiento y funcionalidades del sistema (Figura 24):

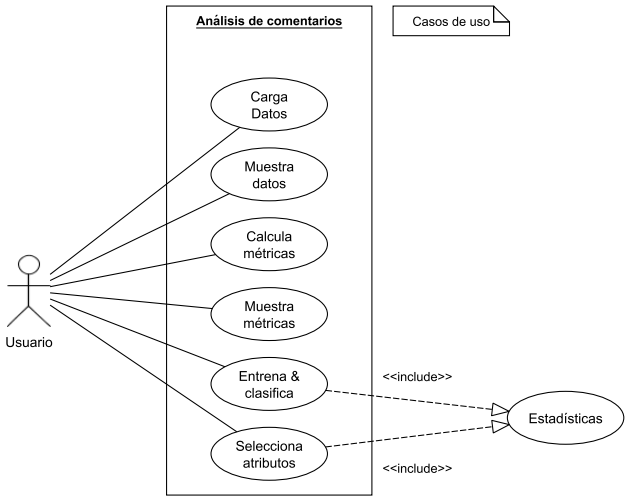


Figura 24: Casos de uso.

1. **CARGA DE DATOS:** El usuario carga los comentarios que desea analizar y que se encuentran almacenados en la base de datos del corpus HOpinion. Estos textos han sido anotados previamente con información morfológica mediante *Freeling* y trasformados a *xml* mediante *AnCora-Pipe* (ver Apartado 3.3.1.1). En AToP se puede escoger entre MySQL o Postgresql como sistema de gestión de bases de datos.

Figura 25: Carga de datos en AToP.

2. **MUESTRA DE DATOS:** AToP cuenta con una interfaz que permite visualizar la información que ha sido cargada desde la base de datos y modificar algunos de los valores que por defecto aparecen en los campos.

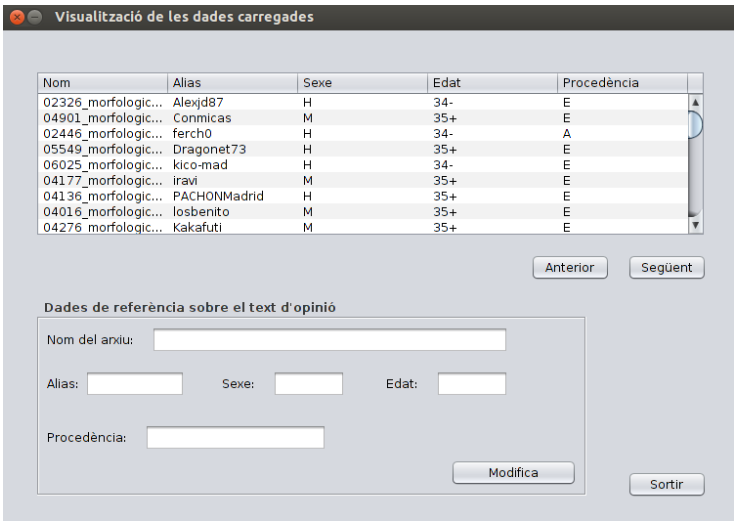


Figura 26: Visualización de datos en AToP.

3. **CÁLCULO DE MÉTRICAS:** AToP permite calcular 32 métricas de riqueza léxica que pueden ser seleccionadas por el usuario de manera individual. Las métricas de riqueza léxica constituyen los atributos con los cuales se caracterizan los comentarios y que sirven como conocimiento en el proceso de aprendizaje automático. AToP utiliza la herramienta *XmlStarlet* para consultar los archivos *xml* y recoger los valores necesarios con los cuales calcular las diferentes métricas de riqueza léxica.

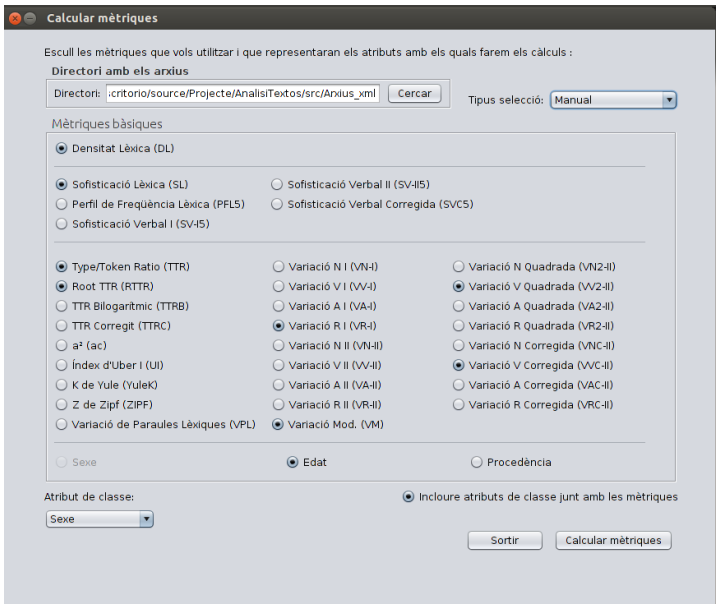


Figura 27: Visualización de las 32 métricas disponibles en AToP.

4. **MUESTRA DE MÉTRICAS:** AToP cuenta con una interfaz que permite visualizar los resultados de las métricas calculadas para cada uno de los comentarios.

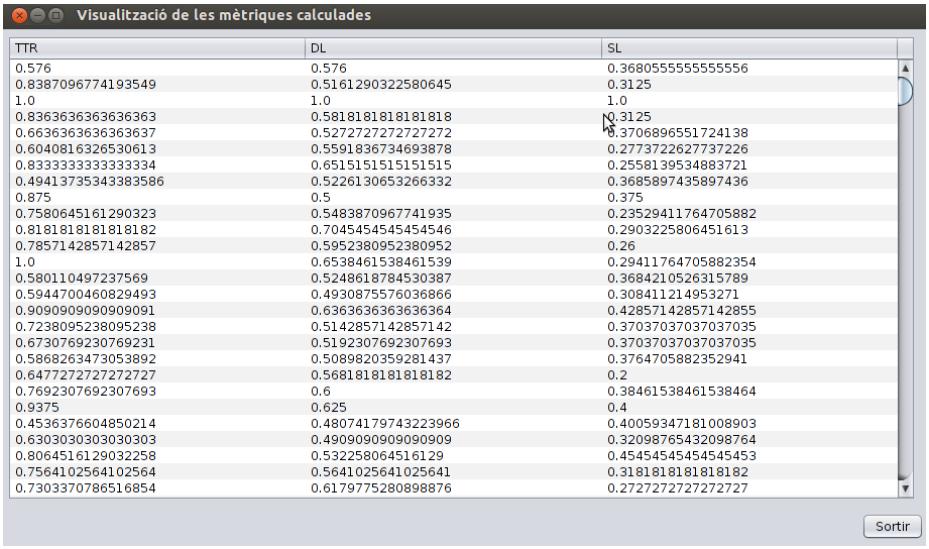


Figura 28: Resultado del cálculo de tres métricas en AToP.

5. **PRUEBA DE ENTRENAMIENTO Y CLASIFICACIÓN:** El usuario puede seleccionar las funciones de clasificación, el número de *folds* para la validació cruzada, la lista de atributos y el conjunto de entrenamiento con los que quiere hacer las pruebas. AToP basa su funcionamiento en Weka[6] del que invoca las clases y componentes necesarios para efectuar las pruebas de clasificación. La lista completa de algoritmos disponibles en Weka (y por lo tanto en AToP) se detalla en el Capítulo 3 (Cuadro 15).

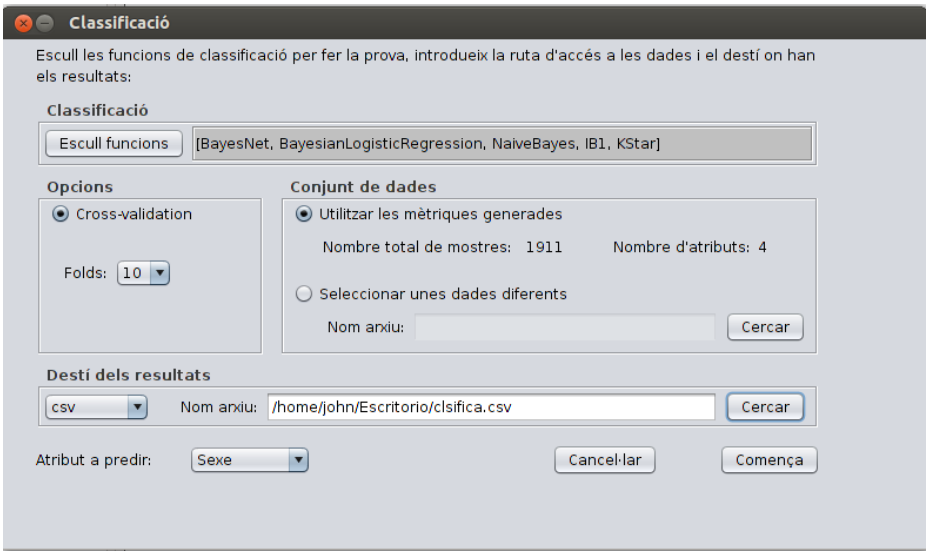


Figura 29: Selección de las funciones de clasificación en AToP.

6 http://www.cs.waikato.ac.nz/ml/weka/

6. **PRUEBA DE SELECCIÓN DE ATRIBUTOS:** El usuario puede seleccionar los métodos de selección y búsqueda con los que quiere hacer las pruebas. La lista completa de los métodos disponibles en Weka (y por lo tanto en AToP) para realizar las pruebas de selección se detalla en el Capítulo 3 (Cuadro 16).

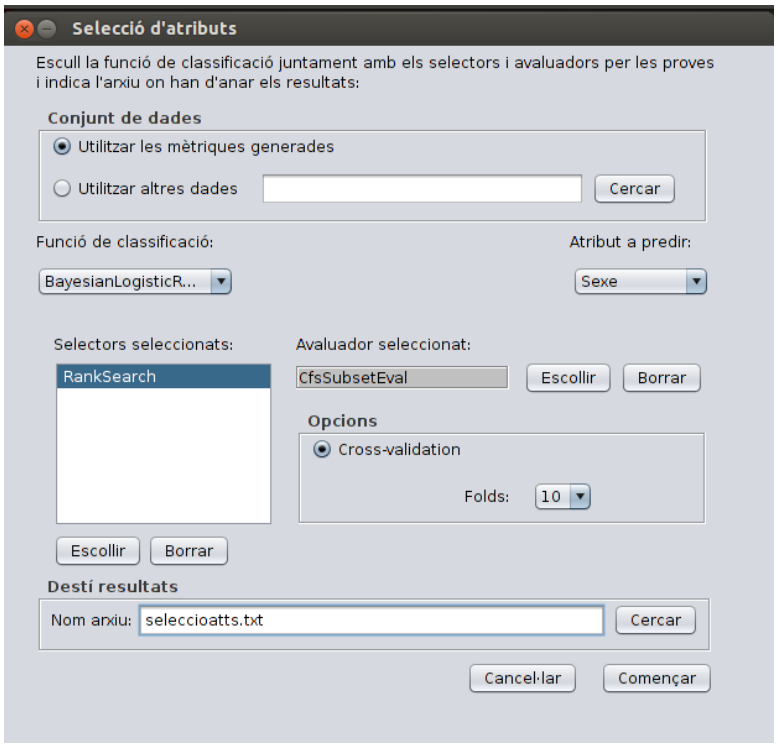


Figura 30: Selección de atributos en AToP.

7. **ESTADÍSTICAS:** AToP genera un fichero en formato CSV como resultado de las pruebas de clasificación donde se guardan los 10 porcentajes de acierto del *cross-validation* para cada uno de los clasificadores seleccionados.

Conjunt de dades: Mètriques										
Atributs:	1.TTR	2.DL	3.SL							
Funció classificació	fold 1	fold 2	fold 3	fold 4	fold 5	fold 6	fold 7	fold 8	fold 9	fold 10
KStar	49.48	54.97	52.88	51.31	50.79	57.59	57.07	51.83	56.54	57.07
NaiveBayes	54.17	55.50	59.16	56.02	58.64	55.50	54.97	52.36	60.21	53.40
BayesianLogisticRegression	53.65	55.50	58.12	54.45	59.16	57.07	55.50	52.88	62.83	58.12
BayesNet	54.17	55.50	59.16	56.02	58.64	55.50	54.97	52.36	60.21	53.40
IB1	50.00	53.93	54.45	53.40	49.21	50.79	51.83	47.64	50.26	52.36

Figura 31: Ejemplo de las estadísticas que proporciona AToP.

La herramienta AToP se puede descargar desde la página web del Centre de Llenguatge i Computació[7] junto con la base de datos del corpus HOpinion y el trabajo de fin de curso de Bakary Singateh [Singateh, 2013] en el que se detalla el proceso de implementación y puesta en funcionamiento de esta herramienta.

7 http://clic.ub.edu/

BIBLIOGRAFÍA

Abney, Steven

1996 «Part-of-Speech Tagging and Partial Parsing», en *Corpus-Based Methods in Language and Speech*, Kluwer Academic Publishers, pág. 118-136.

Adam, Jean-Michel

2011 *Les textes: types et prototypes: Récit, description, argumentation, explication et dialogue*, Linguistique, Armand Colin.

Alba, Joseph y Wesley Hutchinson

1987 «Dimensions of consumer expertise», *Journal of Consumer Research*, 13, 4, pág. 411-454.

Amiri, Hadi y Tat-Seng Chua

2012 *Sentiment Classification Using the Meaning of Words*, inf. téc., Association for the Advancement of Artificial Intelligence.

Aparicio, Juan, Mariona Taulé y M. Antònia Martí

2008 «AnCora-Verb: A Lexical Resource for the Semantic Annotation of Corpora», en *Proceedings of 6th International Conference on Language Resources and Evaluation*, Marrakesh (Morocco).

Aue, Anthony y Michael Gamon

2005 «Customizing Sentiment Classifiers to New Domains: a Case Study», en *International Conference on Recent Advances in Natural Language Processing, RANLP-05*, Borovets, BG.

Bakliwal, Akshat, Piyush Arora, Ankit Patil y Vasudeva Varma

2011 «Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)», en Asian Federation of Natural Language Processing, Chiang Mai, Thailand, cap. Towards Enhanced Opinion Classification using NLP Techniques, pág. 101-107.

Barrera, Luis

1995 *Discurso y Literatura*, Caracas: La Casa de Bello.

Barzilay, Regina y Mirella Lapata

2005 «Modeling Local Coherence: An Entity-based Approach», en *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, Association for Computational Linguistics, Ann Arbor, Michigan, pág. 141-148.

Bejan, Cosmin

2009 *Learning Event Structures from Text*, University of Texas at Dallas. Graduate Program in Computer Science.

Bekkerman, Ron y James Allan

2003 *Using Bigrams in Text Categorization*.

Bertran, Manuel, Oriol Borrega, Mariona Taulé y M. Antònia Martí

2010 *AnCoraPipe: A new tool for corpora annotation*, Barcelona, España.

Biber, Douglas

1992 «The multi-dimensional approach to linguistic analyses of genre variation: An overview of methodology and findings», *Computers and the Humanities*, 26, 5-6, pág. 331-345.

Bieler, Heike, Stefanie Dipper y Manfred Stede

2007 «Identifying Formal and Functional Zones in Film Reviews», en *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pág. 75-78.

Botana, Guillermo, Ricardo Olmos, José León y Francisco Molinero

2008 «Variantes a la extracción automática de vecinos semánticos mediante LSA y al algoritmo de predicación», en SEPEX 08, 7 Congreso de la Sociedad Española de Psicologáa Experimental, San Sebastián, España, pág. 1-11.

Breeder, P., G. Extra y R. van Hout

1986 «Measuring lexical richness and diversity in second language research», *Polyglot*, 8, pág. 1-16.

Breiman, Leo, Jerome Friedman, Charles Stone y R.A. Olshen

1984 *Classification and Regression Trees*, The Wadsworth and Brooks-Cole statistics-probability series, Taylor & Francis, ISBN: 9780412048418.

Brill, Eric

1994 «Some Advances in Transformation-Based Part of Speech Tagging», en *In Proceedings of the twelfth national conference on artificial intelligence*, pág. 722-727.

Buczynski, Aleksander y Aleksander Wawer

2008 «Automated classification of product review sentiments in Polish», *Intelligent Information Systems*, pág. 213-217.

Cambria, Erik, Bjorn Schuller, Yunqing Xia y Catherine Havasi

2013 «New Avenues in Opinion Mining and Sentiment Analysis», *IEEE Intelligent Systems*, 28, 2, pág. 15-21.

Carreiras, Manuel F.

1992 «Estrategias de análisis sintáctico en el procesamiento de frases: cierre temprano vs. cierre tardío», *Cognitiva*, 4, 1, pág. 3-28.

Carroll, J.

1964 *Language and thought*, Englewood Cliffs, NJ: Prentice-Hall.

Carroll, Mary y Monique Lambert

2003 «Information Structure and the Dynamics of Language Acquisition», en John Benjamins, Amsterdam, cap. Information structure in narratives and the role of grammaticised knowledge: A study of adult French and German learners of English, pág. 267-287.

Chambers, Nathanael

2011 *Inducing Event Schemas and their Participants from Unlabeled Text*, Tesis doct., PhD Dissertation, Stanford University.

Chambers, Nathanael y Dan Jurafsky

2008a «Jointly combining implicit constraints improves temporal ordering», en *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, Association for Computational Linguistics, Honolulu, Hawaii, pág. 698-706.

2008b «Unsupervised Learning of Narrative Event Chains», en *Proceedings of ACL-08: HLT*, Association for Computational Linguistics, Columbus, Ohio, pág. 789-797.

2009 «Unsupervised learning of narrative schemas and their participants», en *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, Association for Computational Linguistics, Suntec, Singapore, vol. 2, pág. 602-610, ISBN: 978-1-932432-46-6.

2010 *A Database of Narrative Schemas*.

Chardon, Baptiste, Farah Benamara, Yannick Mathieu, Vladimir Popescu y Nicholas Asher

2013 «Measuring the Effect of Discourse Structure on Sentiment Analysis», en *Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Text Processing - Volume 2*, CICLing'13, Springer-Verlag, Samos, Greece, pág. 25-37.

Chaudron, C. y K. Parker

1990 «Discourse markedness and structural markedness: The acquisition of English noun phrases», *Studies in Second Language Acquisition*, 12, pág. 43-64.

Chenlo, José M., Alexander Hogenboom y David E. Losada

2013 «Sentiment-Based Ranking of Blog Posts Using Rhetorical Structure Theory», en *Natural Language Processing and Information Systems - 18th International Conference on Applications of Natural Language to Information Systems, NLDB 2013, Salford, UK*, Lecture Notes in Computer Science, Springer, vol. 7934, pág. 13-24.

2014 «Rhetorical Structure Theory for polarity estimation: An experimental study», *Data & Knowledge Engineering*, 94, Part B, pág. 135-147.

Cohen, Jacob

1960 «A coefficient of agreement for nominal scales», *Educational and Psychological Measurement*, 20, pág. 37-46.

Comelles, Elisabet, Victoria Arranz e Irene Castelln

2010 «Constituency and Dependency Parsers Evaluation», *Procesamiento del Lenguaje Natural*, 45, pág. 59-66.

Cruz Mata, Ferm´n L.

2012 *Extracción de opiniones sobre caracter´sticas: un enfoque práctico adaptable al dominio*, Colección de monograf´as de la Sociedad Española para el Procesamiento del Lenguaje Natural, Sociedad Española para el Procesamiento del Lenguaje Natural.

Crystal, D.

1982 *Profiling Linguistic Disability*, London: Edward Arnold.

Cui, Hang, Vibhu Mittal y Mayur Datar

2006 «Comparative Experiments on Sentiment Classification for Online Product Reviews», en *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*, AAAI'06, AAAI Press, Boston, Massachusetts, pág. 1265-1270, ISBN: 978-1-57735-281-5.

Dave, Kushal, Steve Lawrence y David Pennock

2003 «Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews», en *Proceedings of the 12th International World Wide Web Conference (WWW2003)*, Budapest, Hungary, pág. 519-528.

De Europa, Consejo

2002 *Marco común de referencia para las lenguas: aprendizaje, enseñanza y evaluación*.

Ding, Xiaowen y Bing Liu

2010 «Resolving object and attribute coreference in opinion mining», en *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, Association for Computational Linguistics, Beijing, China, pág. 268-276.

Dugast, D.

1979 *Vocabulaire et stylistique. I Thtre et Dialogue. Travaux de linguistique quantitative*, Geneva: Slatkine-Champion.

Ebrahim, Neda, Mohammad Fathian y Reza Gholamian

2012 «Sentiment Classification of Online Product Reviews Using Product Features», en *International Journal of Information Processing and Management(IJIPM)*, 3, vol. 3, pág. 30-35.

Eggins, Suzanne

2004 *An introduction to systemic functional linguistics*, Continuum International Publishing Group, London.

Engber, C.

1995 «The relationship of lexical proficiency to the quality of ESL compositions», *Journal of Second Language Writing*, 4, 4, pág. 139-155.

Farra, Noura, Elie Challita, Rawad Abou Assi y Hazem Hajj

2010 «Sentence-Level and Document-Level Sentiment Mining for Arabic Texts», en *Proceedings of the 2010 IEEE International Conference on Data Mining Workshops*, ICDMW '10, IEEE Computer Society, Washington, DC, USA, pág. 1114-1119.

Feng, Vanessa y Graeme Hirst

2014 «A Linear-Time Bottom-Up Discourse Parser with Constraints and Post-Editing», en *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Baltimore, Maryland, pág. 511-521.

Fernández, Javi, José Gómez y Patricio Martínez-Barco

2014a «A Supervised Approach for Sentiment Analysis using skipgrams», en *Proceedings of the Workshop on Natural Language Processing in the 5th Information Systems Research Working Days (JISIC)*, Association for Computational Linguistics, pág. 30-36.

2014b «Análisis de sentimientos multilingüe en la Web 2.0», en *Proc. IV Jornadas PLN-TIMM*, Cazalla de la Sierra, España, pág. 19-21.

Finn, Aidan y Nicholas Kushmerick

2003 «Learning to Classify Documents According to Genre», en *In IJCAI-03 Workshop on Computational Approaches to Style Analysis and Synthesis*.

2006 «Learning to Classify Documents According to Genre: Special Topic Section on Computational Analysis of Style», *Journal of the American Society for Information Science and Technology*, 57, 11, pág. 1506-1518.

Frazier, Lyn

1985 «Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives», en Cambridge University Press, Cambridge, UK, cap. Syntactic complexity, pág. 129-189.

Frazier, Lyn y Charles Clifton

1998 «Reanalysis in Sentence Processing», en Dordrecht: Kluwer Academic Publishers, Cambridge, UK, cap. Sentence Reanalysis, and Visibility, pág. 143-176.

Freund, Y. y L. Mason

1999 «The alternating decision tree learning algorithm», *ICML*, pág. 124-133.

Ganu, Gayatree, Amélie Marian y Noémie Alhadad

2010 *URSA - User Review Structure Analysis: Understanding Online Reviewing Trends*, inf. téc., Rutgers DCS Technical Report No. 668.

Garg, Navneet

2013 «Movie Review Mining in Punjabi», *International Journal of Application or Innovation in Engineering & Management (IJAIEM)*, 2, 12, pág. 372-375.

Ghorbel, Hatem y David Jacot

2011 «Sentiment Analysis of French Movie Reviews», en *Advances in Distributed Agent-Based Retrieval Tools*, pág. 97-108.

Gibson, Edward

1998 «Linguistic complexity: locality of syntactic dependencies», *Cognition*, 68, 1, pág. 1-76.

2000 «Image, language, brain», en Cambridge University Press, MA: MIT Press, Cambridge, UK, cap. The dependency locality theory: a distance-based theory of linguistic complexity, pág. 95-126.

Goldberg, Andrew, Nathanael Fillmore, David Andrzejewski, Zhiting Xu, Bryan Gibson y Xiaojin Zhu

2009 «May all your wishes come true: a study of wishes and how to recognize them», en *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, Association for Computational Linguistics, Boulder, Colorado, pág. 263-271.

Granger, S. y M. Wynne

2000 «Corpora galore: Analyses and techniques in describing English», en Amsterdam: Rodopi, cap. Optimising measures of lexical variation in EFL learner corpora.

Graovac, Jelena y Gordana Pavlovic-Lazetic

2014 «Language-Independent Sentiment Polarity Detection in Movie Reviews: A Case Study of English and Spanish», *ICT Innovations 2014 Web Proceedings*, pág. 13-22.

Grefenstette, Gregory, Yan Qu, James Shanahan y David Evans

2004 «Coupling niche browsers and affect analysis for an opinion mining application», en *In Proceedings of Recherche dï¿Information Assiste par Ordinateur (RIAO*.

Grosse, E.

1976 *Text und Kommunikation*, Köln, Mainz, Stuttgart, Berlin.

Guiraud, P.

1960 *Problemes et methodes de la statistique linguistique*, Dordrecht, The Netherlands: D. Reidel.

Guthrie, David, Ben Allison, Wei Liu, Louise Guthrie y Yorick Wilks

2006 «A Closer Look at Skip-gram Modelling», en *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, European Language Resources Association (ELRA), Genoa, Italy.

Hajishirzi, Hannaneh, Julia Hockenmaier, Erik T. Mueller y Eyal Amir

2011 «Reasoning about RoboCup Soccer Narratives», en *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, ed. por Fabio Gagliardi Cozman y Avi Pfeffer, AUAI Press, pág. 291-300.

Hajishirzi, Hannaneh y Erik T. Mueller

2012 «Question Answering in Natural Language Narratives Using Symbolic Probabilistic Reasoning», en *Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference*, ed. por G. Michael Youngblood y Philip M. McCarthy, AAAI Press, pág. 38-43, ISBN: 978-1-57735-558-8.

Hall, Mark y Eibe Frank

2008 «Combining Naive Bayes and Decision Tables», en *Proceedings of the 21st Florida Artificial Intelligence Society Conference (FLAIRS)*, AAAI press, pág. 318-319.

Halliday, Michael

1978 *Language as social semiotic: The social interpretation of language and meaning*, University Park Press, Maryland.

«Language, context, and text: aspects of language in a social-semiotic perspective» 1985 , en ed. por Michael Halliday y Ruqaiya Hasan, Deakin University Press, Australia, cap. The texture of a text.

Harley, B. y M. King

1989 «Verb lexis in the written compositions of young L2 learners», *Studies in Second Language Acquisition*, 11, pág. 415-440.

Hasan, Ruqaiya

1996 «Ways of saying: ways of meaning. Selected papers of Ruqaiya Hasan», en ed. por Carmel Cloran, David Butt y Geoff Williams, London ; New York : Cassell, cap. The nursery tale as genre.

Hatim, B. e I. Mason

1990 *Discourse and the translator*, Language in social life series, Longman.

Heerschop, Bas, Frank Goossen, Alexander Hogenboom, Flavius Frasincar, Uzay Kaymak y Franciska de Jong

2011 «Polarity Analysis of Texts Using Discourse Structure», en *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, ACM, Glasgow, Scotland, UK, pág. 1061-1070.

Herdan, G.

1960 *Quantitative linguistics*, Butterworth, London.

Hyltenstam, K.

1988 «Lexical characteristics of near-native second-language learners of Swedish», *Journal of Multilingual and Multicultural Development*, 9, pág. 67-84.

Isenberg, Horst

1976 «Einige Grundbegriffe für eine linguistische Texttheorie», en *Probleme der Textgrammatik*, ed. por Danes & Viehweger, Akademie der Wissenschaften, Berlín, pág. 47-145.

Jalilevand, Nahid y Mona Ebrahimipour

2014 «Three measures often used in language samples analysis», *Journal of Child Language Acquisition and Development - JCLAD*, 2, 1, pág. 1-12.

Jarvis, S.

2002 «Short texts, best-fitting curves and new measures of lexical diversity», *Language Testing*, 19, 1, pág. 57-84.

Joshi, Mahesh y Carolyn Penstein

2009 «Generalizing Dependency Features for Opinion Mining», en *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, ACLShort '09, Association for Computational Linguistics, Suntec, Singapore, pág. 313-316.

Joty, Shafiq, Giuseppe Carenini, Raymond Ng y Yashar Mehdad

2013 «Combining Intra- and Multi-sentential Rhetorical Parsing for Document-level Discourse Analysis.», en *ACL (1)*, The Association for Computer Linguistics, pág. 486-496.

Kemper, Susan, Lesa Hoffman, RaLynn Schmalzried, Ruth Herman y Doug Kieweg

2011 «Tracking Talking: Dual Task Costs of Planning and Producing Speech for Young versus Older Adults», *Aging, Neuropsychology, and Cognition*, 18, pág. 257-279.

Kessler, Jason S. y Nicolas Nicolov

2009 «Targeting Sentiment Expressions through Supervised Ranking of Linguistic Configurations», en *ICWSM*, ed. por Eytan Adar, Matthew Hurst, Tim Finin, Natalie S. Glance, Nicolas Nicolov y Belle L. Tseng, The AAAI Press.

Klein, Dan y Christopher Manning

2003 «Accurate Unlexicalized Parsing», en *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, Association for Computational Linguistics, Sapporo, Japan, pág. 423-430.

Kouloumpis, Efthymios, Theresa Wilson y Johanna Moore

2011 «Twitter Sentiment Analysis: The Good the Bad and the OMG!», en *ICWSM*, ed. por Lada A. Adamic, Ricardo A. Baeza-Yates y Scott Counts, The AAAI Press.

Krämer, Nicole y Masashi Sugiyama

2011 «The Degrees of Freedom of Partial Least Squares Regression», *Journal of the American Statistical Association*, 106, 494, pág. 697-705.

Labov, William

1972 «The Transformation of Experience in Narrative Syntax», en *Language in the Inner City*, ed. por William. Labov, University of Pennsylvania Press, Philadelphia, pág. 354-396.

Lascarides, Alex y Nicholas Asher

2007 «Segmented Discourse Representation Theory: Dynamic Semantics with Discourse Structure», en *Computing Meaning: Volume 3*, ed. por H. Bunt y R. Muskens, Kluwer Academic Publishers, pág. 87-124.

Laufer, B. y P. Nation

1995 «Vocabulary size and use: lexical richness in L2 written production», *Applied Linguistics*, 16, pág. 307-322.

Lebret, Remi y Ronan Collobert

2015 «N-gram-based low-dimensional representation for document classification», *Under review as a conference paper at ICLR 2015*.

Lee, Yong-Bae y Sung Hyon Myaeng

2002 «Text genre classification with genre-revealing and subject-revealing features», en *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '02, ACM, Tampere, Finland, pág. 145-150.

Li, Boyang, Darren Appling, Stephen Lee-Urban y Mark Riedl

2012 *Learning Sociocultural Knowledge via Crowdsourced Examples*.

Li, Boyang, Stephen Lee-Urban y Mark O. Riedl

2012 «Toward Autonomous Crowd-Powered Creation of Interactive Narratives», en *Intelligent Narrative Technologies 5, Papers from the 2012 AIIDE Workshop*, pág. 20-25.

Li, Hao, Yu Chen, Heng Ji, Smaranda Muresan y Dequan Zheng

2012 «Combining Social Cognitive Theories with Linguistic Features for Multi-genre Sentiment Analysis», en *Proceedings of the Pacific Asia Conference on Language, Information and Computation (PACLIC2012)*, pág. 127-136.

Li, Shoushan, Sophia Lee, Ying Chen, Chu-Ren Huang y Guodong Zhou

2010 «Sentiment Classification and Polarity Shifting», en *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, Association for Computational Linguistics, Beijing, China, pág. 635-643.

Lin, C., Y. He, R. Everson y S. Ruger

2012 «Weakly Supervised Joint Sentiment-Topic Detection from Text», *IEEE Transactions on n Knowledge and Data Engineering*, 24, 6, pág. 1134-1145.

Linnarud, M.

1986 *Lexis in composition: A performance analysis of Swedish learners' written English*, Lund: CWK Gleerup.

Liu, Bing

2008 *Opinion Mining*.

Liu, Jingjing y Stephanie Seneff

2009 «Review Sentiment Scoring via a Parse-and-paraphrase Paradigm», en *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, EMNLP '09, Association for Computational Linguistics, Singapore, pág. 161-169.

Loureda Lamas, Óscar

2009 *Introducción a la tipología textual*, Arco Libros, Madrid, España.

Lu, X.

2011 «The relationship of lexical richness to the quality of ESL learners' oral narratives», *The Modern Language Journal*.

Maas, Andrew L., Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng y Christopher Potts

2011 «Learning Word Vectors for Sentiment Analysis», en *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, Association for Computational Linguistics, Portland, Oregon, pág. 142-150.

Maas, H.

1972 «Zusammenhang zwischen wortschatzumfang und länge eines textes», *Zeitschrift für Literaturwissenschaft und Linguistik*, 8, pág. 73-79.

Malvern, D., B. Richards, N. Chipere y P. Duran

2004 *Lexical diversity and language development: Quantification and assessment*, Houndmills.

Mann, William y Sandra Thompson

1988 «Rhetorical structure theory: Toward a functional theory of text organization», *Text*, 8, 3, pág. 243-281.

Manning, Christopher e Hinrich Schutze

1999 *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, MA, USA.

Màrquez, Lluís

2002 «Tratamiento del Lenguaje Natural», en ed. por M. Antònia Martí y Joaquim Llisterri, Edicions Universitat de Barcelona, Barcelona, cap. Aprendizaje automático y procesamiento del lenguaje natural, pág. 133-188.

Martin, James

1992 *English text: System and structure*, John Benjamins.

2001 «Language, register and genre», en *Analysing English in a Global Context: a reader*, ed. por A. Burns y C. Coffin, Teaching English Language Worldwide, Routledge, Clevedon, pág. 149-166.

Martineau, Justin y Tim Finin

2009 «Delta TFIDF: an Improved Feature Space for Text Analysis», en *Proceedings of the Third International ICWSM Conference*, Association for the Advancement of Artificial Intelligence, pág. 258-261.

Martínez de Sousa, Joseé

1995 *Diccionario de lexicografía práctica*, Registro, Vox Biblograf.

Matsumoto, Shotaro, Hiroya Takamura y Manabu Okumura

2005 «Sentiment Classification Using Word Sub-sequences and Dependency Sub-trees», en *Proceedings of PAKDD'05, the 9th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*.

McClure, E.

1991 «A comparison of lexical strategies in L1 and L2 written English narratives», *Pragmatics and Language Learning*, 2, pág. 141-154.

Ménard, Nathan

1983 *Mesure de la richesse lexicale: théorie et vérifications expérimentales: études stylométriques et sociolinguistiques*, Travaux de linguistique quantitative, Slatkine.

Meyers, Adam, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young y Ralph Grishman

2004 «The NomBank Project: An Interim Report», en *In Proceedings of the NAACL/HLT Workshop on Frontiers in Corpus Annotation*.

Minsky, Marvin

1975 «Minsky's Frame System Theory», en *Proceedings of the 1975 Workshop on Theoretical Issues in Natural Language Processing*, TINLAP '75, Association for Computational Linguistics, Cambridge, Massachusetts, pág. 104-116.

Moyle, Maura y Steven Long

2013 «Encyclopedia of Autism Spectrum Disorders», en Springer, cap. Index of Productive Syntax (IPSyn), pág. 1566-1568.

Mukherjee, Subhabrata y Pushpak Bhattacharyya

2012 «Sentiment Analysis in Twitter with Lightweight Discourse Analysis», en *COLING 2012, 24th International Conference on Computational Linguistics*, Mumbai, India, vol. 7181, pág. 1847-1864.

Mullen, Tony y Nigel Collier

2004 «Sentiment analysis using support vector machines with diverse information sources», en *In Proceedings of Conference on Empirical Methods in Natural Language Processing*.

Na, Jin-Cheon, Haiyang Sui, Christopher Khoo, Syin Chan y Yunyun Zhou

2004 «Effectiveness of Simple Linguistic Processing in Automatic Sentiment Classification of Product Reviews», en *Knowledge Organization and the Global Information Society: Proceedings of the Eighth International ISKO Conference*, pág. 49-54.

Na, Jin-Cheon y Tun Thet

2009 «Effectiveness of Web Search Results for Genre and Sentiment Classification», *Journal of Information Science*, 35, 6, pág. 709-726.

Narbona, Antonio

1989 *Sintaxis española: nuevos y viejos enfoques*, Ariel.

Ochs, Elinor

1997 «Discourse Studies: A Multidisciplinary Introduction. Vol. 1 (Discourse as Structure and Process)», en ed. por T.A van Dijk, SAGE, London, cap. Narrative, pág. 184-207.

Ortega, Lourdes

2003 «Syntactic Complexity Measures and their Relationship to L2 Proficiency: A Research Synthesis of College-level L2 Writing», *Applied Linguistics*, 4, 24, pág. 492-518.

Ortiz, Antonio Moreno, Francisco Pineda Castillo y Rodrigo Hidalgo García

2010 «Análisis de Valoraciones de Usuario de Hoteles con Sentitext: un sistema de análisis de sentimiento independiente del dominio», *Procesamiento del Lenguaje Natural*, 45, pág. 31-39.

Padró, Lluís

2011 «Analizadores Multilingües en FreeLing», *Linguamatica*, 3, 2 (dic. de 2011), pág. 13-20.

Pajupuu, Hill, Krista Kerge y Rene Altrov

2012 «Lexicon-based detection of emotion in different types of texts: Preliminary remarks.», en *Estonian Papers in Applied Linguistics*, pág. 171-184.

Pakhomov, Serguei, Dustin Chacon, Mark Wicklund y Jeanette Gundel

2010 «Computerized assessment of syntactic complexity in Alzheimer's disease: a case study of Iris Murdoch's writing», *Behavior Research Methods*, 43, 1, pág. 136-144.

2011 «Computerized assessment of syntactic complexity in Alzheimer's disease: a case study of Iris Murdoch's writing», *Behavior Research Methods*, 43, 1, pág. 136-144.

Palmer, Martha, Daniel Gildea y Paul Kingsbury

2005 «The Proposition Bank: An Annotated Corpus of Semantic Roles», *Comput. Linguist.*, 31, 1, pág. 71-106, ISSN: 0891-2017.

Paltoglou, Georgios y Mike Thelwall

2010 «A Study of Information Retrieval Weighting Schemes for Sentiment Analysis», en *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, Association for Computational Linguistics, Uppsala, Sweden, pág. 1386-1395.

Pang, Bo

2006 *Automatica analysis of document sentiment*, Tesis doct., Faculty of the Graduate School of Cornell University.

Pang, Bo y Lillian Lee

2004 «A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts», en *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Association for Computational Linguistics, Stroudsburg, PA, USA.

2008 *Opinion Mining and Sentiment Analysis: Foundations and Trends in Information Retrieval*, Now Publishers Inc., Hanover, MA, USA.

Pang, Bo, Lillian Lee y Shivakumar Vaithyanathan
2002 «Thumbs up: sentiment classification using machine learning techniques», en *EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, Association for Computational Linguistics, Morristown, NJ, USA, pág. 79-86.

Petitjean, André
1987 «Exercices: analyses et productions de textes descriptifs», *Pratiques: théorie, pratique, pédagogie*, 56, pág. 80-100.

Polanyi, Livia y Annie Zaenen
2005 «Contextual valence shifters», en *Computing Attitude and Affect in Text*, Springer, pág. 1-10.

Pustejovsky, J., P. Hanks, R. Sauri, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro y M. Lazo
2003 «The TIMEBANK Corpus», en *Proceedings of Corpus Linguistics 2003*, Lancaster, pág. 647-656.

Rahate, Rohini y Emmanuel M
2013 «Feature Selection for Sentiment Analysis by using SVM», *International Journal of Computer Applications*, 84, 5, pág. 24-32.

Rauen, Fábio
2009 «Genre in a Changing World», en ed. por Charles Bazerman, Adair Bonini y Débora Figueiredo, Fort Collins, Colorado: The WAC Clearinghouse y Parlor Press, cap. Relevance and Genre: Theoretical and Conceptual Interfaces, pág. 56-76.
2010 «Topics on relevance theory», en EDIPUCRS – Editora Universitária da PUCRS, Porto Alegre, cap. On relevance and irrelevances.

Read, J.
2005 *Assessing vocabulary*, 5.ª ed., Cambridge University Press.

Recasens, Marta
2010 *Coreference: Theory, Annotation, Resolution and Evaluation*, Tesis doct., PhD Dissertation, Universidad de Barcelona.

Reed, Chris y Derek Long
1998 «Generating the Structure of Argument», en *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*, Association for Computational Linguistics, Montreal, Quebec, Canada, pág. 1091-1097.

Regneri, Michaela, Alexander Koller y Manfred Pinkal
2010 «Learning script knowledge with web experiments», en *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, Association for Computational Linguistics, Uppsala, Sweden, pág. 979-988.

Reinhart, Tanya
1984 «Principles of gestalt perception in the temporal organization of narrative text», *Linguistics*, 22, pág. 779-809.

Ricci y Wietsma
2006 «Product Reviews in Travel Decision Making», en *Information and Communication Technologies in Tourism 2006*, pág. 296-307.

Roberto, John, M. Antònia Martí y Paolo Rosso

2011 «Sistemas de Recomendación basados en Lenguaje Natural: Opiniones vs. Valoraciones», *Actas IV Jornadas Tratamiento de la Información Multilingüe y Multimodal (TIMM)*, pág. 45-48.

Roberto, John, Maria Salamó y M. Antònia Martí

2012 «Análisis de la riqueza léxica en el contexto de la clasificación de atributos demográficos latentes», *Procesamiento de Lenguaje Natural*, 1, 48, pág. 97-104.

2013 «Clasificación automática del registro lingüístico en textos del español: un análisis contrastivo», *Linguamática*, 5, 1, pág. 59-67.

2014 «The function of narrative chains in the polarity classification of reviews», *Procesamiento del Lenguaje Natural*, 52, pág. 69-76.

2015a «Genre-Based Stages Classification for Polarity Analysis», en *The 28th Florida Artificial Intelligence Society Conference (FLAIRS), USA*, vol. 1, pág. 1-6.

2015b «Polarity analysis of reviews based on the omission of asymetric sentences», *Procesamiento del Lenguaje Natural*, 54, pág. 77-84.

Rothfels, John y Julie Tibshirani

2010 *Unsupervised sentiment classification of English movie reviews using automatic selection of positive and negative sentiment items*.

Roulet, Eddy

1991 «Une approche discoursive de l'hétérogénéité discoursive», *Etudes des Linguistique Appliquée*, 83, pág. 117-130.

Ruiz Sánchez, Roberto

2005 *Selección de Atributos mediante proyecciones*, Tesis doct., PhD Dissertation, Universidad de Sevilla.

Saleh, Rushdi, M. T. Martín-Valdivia, A. Montejo-Ráez y L. A. Ureña-LóPez

2011 «Experiments with SVM to Classify Opinions in Different Domains», *Expert Syst. Appl.*, 38, 12, pág. 14799-14804, ISSN: 0957-4174.

Schank, Roger C. y Robert P. Abelson

1975 «Scripts, Plans, and Knowledge», en *Proceedings of the 4th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'75, Morgan Kaufmann Publishers Inc., Tblisi, USSR, pág. 151-157.

Schilder, Frank

1997 *Temporal Relations in English and German Narrative Discourse*, Tesis doct., The University of Edinburgh. College of Science, Engineering. School of Informatics. Institute for Communicating y Collaborative Systems.

Sharoff, Serge, Wu Zhili y Markert Katja

2010 «The web library of babel: evaluating genre collections», *Proceedings of Seventh International Conference on Language Resources and Evaluation (LREC10)*, pág. 3063-3070.

Sidorov, Grigori, Sabino Miranda-Jiménez, Francisco Viveros-Jiménez, Alexander Gelbukh, Noé Castro-Sánchez, Francisco Velásquez, Ismael Díaz-Rangel, Sergio Suárez-Guerra, Alejandro Treviño y Juan Gordon

2012 *Empirical Study of Opinion Mining in Spanish Tweets*.

2006 , en *Aprendizaje Automático: conceptos básicos y avanzados*, ed. por Basilio Sierra, Pearson Educación.

Silveira, Marcirio, Larissa de Freitas y Renata Vieira

2012 (eds.), *Hontology: A Multilingual Ontology for the Accommodation Sector in the Tourism Industry*, SciTePress, pág. 149-154.

Singateh, Bakary

2013 *Plataforma en Java per l'Anàlisi de Textos d'Opinió en llenguatge natural (ATOp)*, Tesis de lic., Universitat de Barcelona. Facultat de Matemàtiques.

Smith, J. y C. Kelly

2002 «Stylistic Constancy and Change Across Literary Corpora: Using Measures of Lexical Richness to Date Works», *Computers and the Humanities*, 36, pág. 411-430.

Sperber, Dan y Deirdre Wilson

1986 *Relevance: Communication and Cognition*, Harvard University Press, Cambridge, MA, USA.

Staab, Steffen y Rudi Studer

2009 *Handbook on Ontologies*, 2nd, Springer Publishing Company, Incorporated.

Stenstrom, Anna-Brita

1994 *An introduction to spoken interaction*, Longman, London.

Stone, Philip, Dexter Dunphy, Marshall Smith y Daniel Ogilvie

1966 *The General Inquirer: A Computer Approach to Content Analysis*, ed. por M. I. T. Press, MIT Press.

Swales, John

1990 *Genre Analysis: English in Academic and Research Settings*, Cambridge University Press.

Taboada, Maite

2011 «Stages in an online review genre», *Text and Talk. An Interdisciplinary Journal of Language, Discourse & Communication Studies*, 31, 2, pág. 247-269.

Taboada, Maite, Julian Brooke y Manfred Stede

2009 «Genre-based paragraph classification for sentiment analysis», en *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '09, Association for Computational Linguistics, London, United Kingdom, pág. 62-70.

Taboada, Maite, Julian Brooke, Milan Tofiloski, Kimberly Voll y Manfred Stede

2011 «Lexicon-based Methods for Sentiment Analysis», *Computational Linguistics*, 37, 2, pág. 267-307.

Taboada, Maite y Jack Grieve

2004 «Analyzing Appraisal Automatically», en *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, Stanford, US, pág. 158-161.

Taboada, Maite y William Mann

2006 «Rhetorical Structure Theory: looking back and moving ahead», *Discourse Studies*, 8, pág. 423-459.

Tan, Songbo, Xueqi Cheng, Yuefen Wang y Hongbo Xu

2009 «Adapting Naive Bayes to Domain Adaptation for Sentiment Analysis», *Advances in Information Retrieval*, pág. 337-349.

Tannen, D.

1989 «Talking Voices. Repetition, dialogue and imagery in conversational discourse», en ed. por D. Tannen, required, Cambridge, CUP, cap. Repetition in Conversation: Towards a poetics of Talk.

Taulé, Mariona, M. Antònia Martí y Marta Recasens

2008 «AnCora: Multilevel Annotated Corpora for Catalan and Spanish», en *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, ed. por Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis y Daniel Tapias, European Language Resources Association (ELRA), Marrakech, Morocco.

Templin, M.

1957 *Certain language skills in children: Their development and interrelationships*, Minneapolis: The University of Minnesota Press.

Thordardottir, E. y S. Weismer

2001 «High-frequency verbs and verb diversity in the spontaneous speech of school-age children with specific language impairment», *International Journal of Language and Communication Disorders*, 36, pág. 221-244.

Tong, Richard

2001 «An Operational System for Detecting and Tracking Opinions in On-line Discussions», en *Working Notes of the SIGIR Workshop on Operational Text Classification*, New Orleans, Louisianna, pág. 1-6.

Tu, Zhaopeng, Yifan He, Jennifer Foster, Josef van Genabith, Qun Liu y Shouxun Lin

2012 «Identifying High-Impact Sub-Structures for Convolution Kernels in Document-level Sentiment Classification», en *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, The Association for Computer Linguistics, pág. 338-343.

Turney, Peter

2002 «Thumbs Up or Thumbs Down Semantic Orientation Applied to Unsupervised Classification of Reviews», en *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*, Philadelphia, Pennsylvania, pág. 417-424.

Tweedie, F. y H. Baayen

1998 «How Variable May a Constant be Measures of Lexical Richness in Perspective», *Computers and the Humanities*, 32, pág. 323-352.

Ure, J.

1971 «Talking about text», en University of Birmingham, English Language Research, pág. 27-48.

Verschueren, J.

1999 *Understanding pragmatics*, London: Arnold.

Vilares, David, Miguel A. Alonso y Carlos Gómez-Rodríguez

2013 «Clasificación de polaridad en textos con opiniones en español mediante análisis sintáctico de dependencias», *Procesamiento del Lenguaje Natural*, 50, pág. 13-20.

Vinodhini, G. y R. M. Chandrasekaran

2012 «Sentiment Analysis and Opinion Mining: A Survey», *International Journal of Advanced Research in Computer Science and Software Engineering*, 2, 6, pág. 282-292.

Voll, Kimberly y Maite Taboada

2007 «Not all words are created equal: Extracting semantic orientation as a function of adjective relevance», en *Proceedings of the 20th Australian Joint Conference on Artificial Intelligence*, pág. 337-346.

Wang, Yuan, Zhaohui Li, Jie Liu, Zhicheng He, Yalou Huang y Dong Li

2014 «Word Vector Modeling for Sentiment Analysis of Product Reviews», *NLPCC 2014*, pág. 168-180.

Webber, Bonnie, Markus Egg y Valia Kordoni

2011 «Discourse Structure and Language Technology», *Natural Language Engineering*, 18, 4, pág. 437-490, ISSN: 1351-3249.

2012 «Discourse Structure and Language Technology», *Journal of Natural Language Engineering*, 01, ed. por Ruslan Mitkov, pág. 1-40.

Werlich, E.

1976 *A Text Grammar of English*, UTB Anglistik, Quelle & Meyer.

Wiebe, Janyce, Theresa Wilson, Rebecca Bruce, Matthew Bell y Melanie Martin

2004 «Learning Subjective Language», *Computational Linguistics*, 30, 3, pág. 277-308.

Wilson, Deirdre y Dan Sperber

2004 «The Handbook of Pragmatics», en ed. por Laurence Horn y Gregory Ward, Blackwell Publishing, cap. Relevance theory, pág. 607-632.

Witten, Ian, Eibe Frank, Len Trigg, Mark Hall, Geoffrey Holmes y Sally Cunningham

1999 *Weka: Practical Machine Learning Tools and Techniques with Java Implementations*.

Wolfe-Quintero, K., S. Inagaki y H. Kim

1998 *Second language development in writing: Measures of fluency, accuracy, and complexity*, inf. téc., Honolulu: University of Hawai'i, Second Language Teaching y Curriculum Center.

Wu, Yuanbin, Qi Zhang, Xuanjing Huang y Lide Wu

2009 «Phrase Dependency Parsing for Opinion Mining», en *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, EMNLP '09, Association for Computational Linguistics, Singapore, pág. 1533-1541.

Xie, Rui, Chunping Li, Qiang Ding y Li Li

2014 «Integrating Topic, Sentiment and Syntax for Modeling Online Review», *The Sixth International Conference on Information, Process, and Knowledge Management (eKNOW)*, pág. 137-144.

Xu, Liheng, Kang Liu, Siwei Lai y Jun Zhao

2014 «Product Feature Mining: Semantic Clues versus Syntactic Constituents», en *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014*, Baltimore, MD, USA, vol. 1, pág. 336-346.

Yih, Wen-tau, Geoffrey Zweig y John C. Platt

2012 «Polarity Inducing Latent Semantic Analysis», en *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, Association for Computational Linguistics, Jeju Island, Korea, pág. 1212-1222.

Yngve, Victor

1960 «A model and an hypothesis for language structure», *Proceedings of the American Philosophical Society*, 104, 5, pág. 444-466.

Yule, G.

1944 *The statistical study of literary vocabulary*, Cambridge, UK: Cambridge University Press.

Zhang, Changli, Daniel Zeng, Jiexun Li, Fei-Yue Wang y Wanli Zuo

2009 «Sentiment Analysis of Chinese Documents: From Sentence to Document Level», *Journal of the American Society for Information Science and Technology*, 60, 12, pág. 2474-2487.

zhang, Qi, Yuanbin Wu, Tao Li, Mitsunori Ogihara, Joseph Johnson y Xuanjing Huang

2009 «Mining Product Reviews Based on Shallow Dependency Parsing», en *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, ACM, Boston, MA, USA, pág. 726-727.

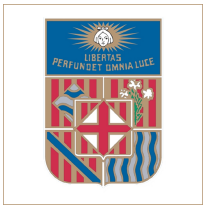
Zhang, Weishi, Kai Zhao, Likun Qiu y Changjian Hu

2009 «SESS: A Self-Supervised and Syntax-Based Method for Sentiment Classification», en *PACLIC*, ed. por Olivia Kwong, City University of Hong Kong Press, pág. 596-605.

Zhang, Yongfeng, Haochen Zhang, Min Zhang, Yiqun Liu y Shaoping Ma

2014 «Do Users Rate or Review: Boost Phrase-level Sentiment Labeling with Review-level Sentiment Classification», en *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, ACM, Gold Coast, Queensland, Australia, pág. 1027-1030.

ÍNDICE ALFABÉTICO


LIBERTAS
PERFUNDET OMNIA LUCE