Análisis de Valoraciones de Usuario de Hoteles con *Sentitext**: un sistema de análisis de sentimiento independiente del dominio

Analyzing Hotel Reviews with Sentitext: a domain-independent, sentiment analysis system

Antonio Moreno Ortiz, Francisco Pineda Castillo, Rodrigo Hidalgo García
Facultad de Filosofía y Letras
Universidad de Málaga
Campus de Teatinos
29071 Málaga
{amo,pineda,rodrigo.hidalgo}@uma.es

Resumen: En este trabajo describimos la primera prueba sistemática realizada para evaluar el rendimiento de *Sentitext*, un sistema de análisis de sentimiento aún en fase de desarrollo. A diferencia de la mayoría de los sistemas existentes, el funcionamiento de Sentitext está basado enteramente en conocimiento lingüístico independiente del dominio, empleando para ello bases de datos léxicas de amplia cobertura desarrolladas al efecto, careciendo de algoritmos de aprendizaje y de clasificador en el sentido estricto. Los resultados arrojados por esta primera prueba, para la que hemos empleado textos de críticas de usuarios a hoteles provenientes de la versión en español de *Tripadvisor*, son extremadamente alentadores, dada la alta tasa de acierto en cuanto a polaridad.
Palabras clave: Análisis de sentimiento, minería de opiniones, valoraciones de usuario.

Abstract: This paper describes a first test run of Sentitext, a sentiment analysis system under development. Unlike most existing systems, Sentitext is entirely based on linguistic knowledge and independent of any domain, using a wide coverage lexical database and lacking learning algorithms or classifiers, strictly speaking. Results on this first test, for which a collection of hotel review texts from *Tripadvisor* has been used, are extremely encouraging, given the high polarity hit rate.
Keywords: Sentiment analysis, opinion mining, on-line user reviews.

1 Introducción

Con la expansión de Internet y de sitios Web 2.0 tales como blogs, prensa en línea con participación de lectores y webs de comparativa y evaluación de productos y servicios, entre otros, cada vez son más las fuentes de información y las opiniones de usuarios que están a disposición de todos los usuarios de la Red. Esta demanda de información actualizada y de primera mano se hace más necesaria en el caso de las empresas, que están interesadas en saber cómo son valorados sus productos y servicios por los usuarios finales.

Sin embargo, la mayor parte de esta información se encuentra codificada como texto, es decir, datos sin estructura aparente para la máquina. Por este motivo, desde hace tiempo se ha venido demostrando la necesidad de crear herramientas informáticas de PLN capaces de analizar estas grandes cantidades de texto y extraer y sintetizar información de forma automatizada, surgiendo así disciplinas como la *minería de textos*, también llamada analítica de textos. Uno de los tipos de información codificada en los textos, que ha venido recibiendo cada vez más atención, es lo que se ha dado en llamar el *análisis de sentimiento*, también llamado *minería de opinión* (p. ej., Esuli y Sebastiani, 2006), consistente en la valoración, clasificación del componente axiológico del lenguaje, es decir, aquellos aspectos lingüísticos que codifican la subjetividad, en términos de positividad o

* Esta herramienta ha sido diseñada por el grupo *Tecnolengua* (http://tecnolengua.uma.es) y desarrollada gracias a la financiación de El Jardín de Junio S.L.U. mediante convenio OTRI con la Universidad de Málaga Nº 8.06/5.21.3199-1.

negatividad, del hablante con respecto a aquello de lo que habla.

Pang y Lee (2008) repasan el avance de la investigación en el campo del análisis de sentimiento en la última década, donde el principal foco de atención ha sido el análisis de valoraciones de usuarios de películas (Pang, Lee & Vaithyanathan, 2002), productos electrónicos de consumo (Dave et al., 2003), pero también el análisis del mercado financiero (Das y Chen, 2001), de discursos políticos (Thomas, Pang y Lee, 2006; Miller, D.R., 2004), la prensa mundial (Wiebe et al., 2003; White, 2006; White & Thomson, 2008) o las valoraciones de usuarios en webs de evaluación de productos de carácter general, como *Epinions* (Turney, 2002; Taboada y Grieve, 2004). En español la actividad en el campo es mucho más reducida, aunque hay notables trabajos, como los de Cruz et al. (2008) y Boldrini et al. (2009).

Además de comparar la valoración de Sentitext con la otorgada por el usuario, aportamos un estudio pormenorizado basado en técnicas de análisis del discurso para categorizar las fuentes principales de error y ofrecemos un análisis textual detallado, con el objeto de evaluar el rendimiento actual de la herramienta y el que esperamos alcanzar tras la integración de las reglas de contexto, actualmente en desarrollo.

2 La importancia del dominio

La visión más extendida en el ámbito del análisis del sentimiento es que éste es altamente dependiente del dominio (Aue & Gamon 2005). Es decir, determinadas palabras y unidades fraseológicas que en un dominio determinado serían contempladas como afectivamente neutras pueden adquirir una determinada *polaridad*, u *orientación semántica*, en ese dominio determinado. Por ejemplo, en el ámbito en el que hemos desarrollado este estudio, el de las críticas de hoteles, aspectos como el tamaño o la localización conllevan de forma invariable implicaciones positivas o negativas (todo lo pequeño es negativo, todo lo céntrico o "cerca de" es positivo). Es por ello que los esfuerzos se han enfocado en construir recursos léxicos específicos de dominios determinados, empleando para ello técnicas de adquisición automatizada, con o sin supervisión (Turney y Littman, 2002). Por este motivo se habla de "entrenar" los algoritmos de clasificación. La desventaja obvia de este enfoque es que un clasificador entrenado para un dominio concreto es difícilmente aplicable a otro dominio o textos de opinión. El problema es que existen multitud de dominios especializados, por lo que este enfoque implica el entrenamiento en cada uno de ellos, o bien desarrollar mecanismos de transferencia entre dominios (p. ej., Aue y Gamon, 2005, Tan et al., 2007). Nuestro enfoque es distinto, pues nuestros recursos léxicos están construidos de forma "ciega", es decir, sin tener como objetivo dominio especializado alguno. Lógicamente, esta decisión implica que nuestro analizador no será capaz de identificar segmentos con una determinada orientación semántica en aquellos casos en que dicha orientación esté determinada por el dominio (p. ej., "las camas son minúsculas"), pero, por otro lado, no condiciona nuestro sistema a ningún dominio ni aplicación determinados.

La cuestión fundamental es ¿cuál es la proporción de marcadores afectivos específicos del dominio con respecto a aquellos aplicables a la lengua en general? O, dicho de otra manera, ¿hasta qué punto condicionan estos marcadores de discurso especializado el resultado de la valoración global del texto? En este estudio nos proponemos precisamente cuantificar esta proporción.

3 Sentitext

Sentitext es una aplicación con arquitectura cliente-servidor con un alto nivel de modularidad. La herramienta de análisis principal está desarrollada en C++, para facilitar la integración con el analizador morfológico que utiliza, *Freeling* (Atserias et al., 2006). También hemos desarrollado un cliente basado en Web, desarrollado en Flex, y se utiliza XML como vehículo para el flujo de datos entre el cliente y el servidor. El cliente muestra el resultado del análisis de una forma gráfica de varias maneras, resalta los segmentos con polaridad en el texto original y ofrece varias medidas precisas. Actualmente nos encontramos explorando varias posibilidades en cuanto a las representaciones gráficas de contenidos textuales desde el punto de vista de la afectividad, aspecto en el que no se ha trabajado mucho (Liu et al., 2003).

La información léxica para el análisis afectivo que Sentitext emplea se recupera en tiempo de ejecución de una base de datos

MySQL, manteniendo así una separación estricta entre procesos y datos, y permitiendo dotar de gran flexibilidad e independencia a los encargados de la adquisición del léxico, por un lado, y programadores por otro. De este modo, disponemos de una interfaz Web a la base de datos que facilita corregir errores u omisiones en los lexicones de forma inmediata sobre la marcha. Esta arquitectura modular, con una estricta separación entre procesos y datos, está motivada por la misma naturaleza de la aplicación, que se basa exclusivamente en los datos léxicos almacenados para llevar a cabo la valoración de los textos.

Sentitext se nutre de tres fuentes de datos fundamentales: (i) el léxico de palabras individuales, (ii) el léxico de frases y (iii) las reglas de contexto. Los tres han sido desarrollados de forma manual, pero utilizando herramientas construidas al efecto para facilitar el proceso de adquisición. En el caso del lexicón de palabras individuales se partió utilizando metodologías conocidas como la utilización de un diccionario de sinónimos con palabras semilla (Pitel y Grefenstette, 2008, Mohammad et al., 2009). En nuestro caso, empleamos el diccionario de sinónimos de español de *OpenOffice*, y se creó una interfaz de adquisición que facilitó el proceso, que podríamos calificar de "asistido" más que semiautomático. El lexicón de palabras individuales consta actualmente de unas 10.500 palabras con carga afectiva, marcadas con su polaridad con uno de estos 4 valores: -2 (muy negativa), -1 (negativa), 1 (positiva), 2 (muy positiva). Estas etiquetas fueron asignadas de forma manual y consensuada entre el equipo de lexicógrafos de Tecnolengua, basándose en rasgos semánticos y recurriendo a un corpus de lengua general.

Algo que no ha recibido especial atención en trabajos anteriores, sin embargo, ha sido la inclusión de locuciones o expresiones multi-palabra. Como recuerdan Wilson et al. (2009), a veces se usan palabras positivas en frases negativas (p. ej.: "abuso de confianza") y viceversa, o simplemente un grupo de palabras que de forma individual son neutras conforman una unidad de significado con polaridad (p. ej.: "coser y cantar"). En nuestro caso, decidimos desde un principio adquirir un lexicón de locuciones de amplia cobertura y no específico del dominio. Dicho lexicón contiene en la actualidad más de 17.000 entradas, aunque, a diferencia del lexicón de palabras individuales, no todas tienen carga afectiva. Somos conscientes de que un lexicón de locuciones es mucho más difícil que pueda ser considerado "de gran cobertura", aunque éste es en efecto nuestro objetivo. De hecho, consideramos ambos recursos como no acabados y en continua expansión.

El sistema ha sido diseñado para emplear las mencionadas reglas de contexto, inspiradas en gran medida en la propuesta de Polanyi y Zaenen (2006), en lo que ellas denominan *Context Valence Shifters (CVS)*, o modificadores contextuales de la valencia. Su funcionalidad se basa en describir estos modificadores formalmente, mediante reglas de contexto en las que también se codifica el resultado de la aplicación de la regla. Por ejemplo, una regla puede expresar que la presencia del adverbio "muy" precediendo un adjetivo con carga afectiva intensifica su valencia en el mismo sentido. Hasta la fecha hemos recopilado 350 reglas de contexto, algunas enormemente generales como la del ejemplo anterior, otras muy específicas, y en la actualidad nos encontramos desarrollando diferentes maneras de aplicarlas en el proceso de análisis, por lo que en los resultados que describimos en este trabajo, no se han empleado.

La forma de calcular dicho valor global no es algo trivial. En principio, podría pensarse en una media aritmética de las valencias de las unidades léxicas, pero esto presenta varios problemas, entre ellos, que a un texto de gran longitud con una única palabra positiva se le asignaría el máximo valor, o que a un texto completamente neutro se le asignaría el mismo valor global que a uno con la mitad de palabras positivas y la otra mitad negativas. Actualmente, utilizamos como valor global una media aritmética ponderada, que se somete a una modificación posterior basándose en el índice afectivo: si el texto contiene muchas unidades con valencia afectiva distinta de cero, el valor puede moverse más libremente hacia los extremos, en caso contrario, se tiende a centralizar esta medida.

4 Valoraciones de usuarios de hotel

Tripadvisor es un sitio Web destinado a recoger información turística, sobre todo en lo que respecta a hoteles. Su sistema de recogida de valoraciones de usuarios sigue los patrones estándar: el usuario selecciona un producto

determinado (en este caso un hotel), lo describe según su experiencia con el mismo mediante un texto de longitud variable con un mínimo de 50 caracteres, y agrega una valoración global de 1 a 5, siendo 1 muy negativo y 5 muy positivo. *Tripadvisor* también permite al usuario valorar determinados aspectos del hotel de manera individualizada tales como limpieza, ubicación, habitaciones, servicio, relación calidad-precio y calidad del sueño; sin embargo, es importante resaltar que la valoración global del hotel no se calcula en base a la media de estas subcategorías, sino que el propio usuario otorga la valoración global de forma directa, con independencia de su valoración en las diferentes subcategorías. Este sistema, empleado en otras muchas Webs de *reviews* de productos y servicios, se ha dado en llamar *star rating*, siguiendo el sistema tradicional de clasificación de categorías de hoteles.

Un aspecto que en la literatura sobre análisis de sentimiento se suele dar por hecho es que la valoración vertida por el usuario en el texto se corresponde con la valoración numérica global. Sin embargo, hemos constatado cómo esto no es así en muchas ocasiones. Ofrecemos algunos ejemplos más abajo, donde aportamos los resultados de un estudio pormenorizado de los textos empleando técnicas de análisis del discurso y contrastamos estos resultados con la valoración de *Sentitext*.

4.1 Muestra analizada

Para este trabajo hemos tomado una selección aleatoria de 100 valoraciones. Los únicos criterios de selección fueron los siguientes:

- Similar extensión: el algoritmo de valoración global que emplea Sentitext tiene en cuenta no sólo el número de unidades léxicas con carga afectiva, sino también el número total de palabras.
- Original en español: algunas valoraciones de usuarios incluidas en *tripadvisor.es* son traducciones automáticas de valoraciones en otras lenguas a través de la herramienta Language Weaver.
- Homogeneidad en el objeto de crítica: los hoteles valorados deberían pertenecer a la misma área. Optamos por tomar Londres por ser una de las ciudades con mayor número de valoraciones.
- Distribución proporcional en cuanto al número de críticas para las valoraciones numéricas.

La muestra final tomada, quedó conformada como se resume en la Tabla 1.

Dato	**Valor**
Número de (textos) valoraciones	100
Valoraciones 1 estrella	22
Valoraciones 2 estrellas	18
Valoraciones 3 estrellas	20
Valoraciones 4 estrellas	19
Valoraciones 5 estrellas	21
Número de palabras (media)	133
Palabras léxicas (media)	64

Tabla 1: Datos de la muestra analizada

Un último aspecto que se tuvo en cuenta antes de analizar los textos con Sentitext fueron las faltas ortográficas que ocasionalmente los usuarios pudieran cometer. En la actualidad, es preciso que el *input* de Sentitext sea ortográficamente correcto, de otra manera el lematizador falla y las asignaciones de valencia no son correctas. Por tanto, procedimos a una corrección de los textos, utilizando para ello un corrector ortográfico.

5 Resultados

El rendimiento de Sentitext, como clasificador de la positividad o negatividad de los textos valorativos del tipo que hemos empleado en este trabajo es muy bueno, acercándose al 90% de acierto si no consideramos las valoraciones con 3 estrellas como neutras. En realidad, no deberíamos considerar ninguno de estos textos como "neutros" en el sentido estricto de la palabra, pues todos son inherentemente valorativos. Por esto hemos considerado que la clasificación binaria (+/-) era correcta si la diferencia entre la valoración en número de estrellas otorgada por el usuario y la ofrecida por Sentitext era de 0 ó 1, siendo errónea si la diferencia era de 2 ó más estrellas.

El análisis de los resultados obtenidos lo hemos realizado desde dos puntos de vista. En primer lugar, hemos comparado la valoración del usuario con la obtenida por el analizador. La Tabla 2 muestra un resumen de los resultados obtenidos.

Dato	**Valor**
Coincidencia exacta	37%
Diferencia de 1 estrella	52%
Coincidencia en polaridad	**89%**
Diferencia de 2 estrellas	11%
Diferencia de más de 2 estrellas	0%

Tabla 2: Resultados *star-rating*

Estas cifras nos ofrecen unos resultados muy alentadores, pero sin duda mejorables y el primer paso para mejorarlos es entender el origen de los errores de análisis de nuestra herramienta. A tal fin nuestro siguiente paso ha sido llevar a cabo un estudio pormenorizado del resultado de los análisis en términos de recuperación de información. Sentitext no sólo ofrece una valoración global de los textos, sino que además podemos saber exactamente qué segmentos textuales han sido etiquetados con qué valencia afectiva

En la Tabla 3 mostramos un listado de los resultados obtenidos en notación *star rating*. En cada columna, el formato es "nº de texto: valoración del usuario (vu) – valoración de Sentitext (vs)". En realidad, Sentitext no obtiene originalmente este tipo de resultado, sino que lo calcula a partir de un valor global expresado en porcentaje de afectividad al que denominamos *gValue*, en el que los valores cercanos al 0% serían negativos y los cercanos al 100% serían positivos.

#:vu-vs	#:vu-vs	#:vu-vs	#:vu-vs	#:vu-vs
1: 2-3	21: 4-4	41: 2-3	61: 4-4	81: 1-3
2: 3-4	22: 4-4	42: 1-2	62: 2-4	82: 1-2
3: 3-4	23: 4-4	43: 5-4	63: 3-3	83: 1-3
4: 3-5	24: 4-4	44: 5-5	64: 4-3	84: 3-4
5: 3-3	25: 5-4	45: 5-4	65: 4-4	85: 1-2
6: 5-4	26: 1-3	46: 4-4	66: 4-4	86: 1-3
7: 3-4	27: 5-4	47: 4-4	67: 2-4	87: 1-1
8: 4-4	28: 5-4	48: 5-5	68: 1-2	88: 1-2
9: 4-4	29: 1-1	49: 5-4	69: 2-2	89: 2-3
10: 5-4	30: 4-4	50: 3-3	70: 2-3	90: 5-4
11: 4-4	31: 4-4	51: 3-4	71: 5-4	91: 5-4
12: 2-2	32: 4-4	52: 3-3	72: 3-3	92: 2-4
13: 5-4	33: 4-4	53: 5-4	73: 3-3	93: 5-4
14: 1-2	34: 1-2	54: 3-5	74: 2-3	94: 2-3
15: 5-5	35: 1-1	55: 1-3	75: 2-3	95: 5-4
16: 2-3	36: 3-3	56: 4-4	76: 1-2	96: 5-4
17: 1-2	37: 3-4	57: 2-2	77: 1-1	97: 5-5
18: 1-2	38: 3-3	58: 3-4	78: 2-2	98: 4-4
19: 1-2	39: 3-4	59: 2-3	79: 2-3	99: 2-4
20: 5-4	40: 1-2	60: 1-2	80: 3-3	100: 3-4

Tabla 3: Resultados texto a texto

En la Tabla 3 hemos marcado con un fondo más oscuro, los 11 casos en los que hemos obtenido una diferencia de 2 estrellas, y por tanto consideramos resultados erróneos. Es interesante comprobar cómo la mayoría de estos casos (72,7%) se refieren a valoraciones negativas del usuario, lo que vendría a implicar que Sentitext obtiene mejores resultados con textos positivos.

El siguiente paso para confirmar esta sospecha y obtener datos más concretos es contabilizar y clasificar los segmentos valorativos que realmente aparecen en los textos y compararlos con los resultados del análisis automático. Por *segmento valorativo (SV)* entendemos una secuencia textual que conlleva una opinión o valoración de aquello sobre lo que se habla, tales como "no me gustó nada" o "los baños están muy limpios". Por tanto, no tenemos en cuenta elementos discursivos más abstractos, que se manifiestan en los textos de formas más sutiles, aspectos que discutimos en el apartado 6.3. La Tabla 4 ofrece un resumen de estos datos.

Segmentos	Cantidad	%
SVs totales	**1224**	**100**
Positivos	691	56,45
Negativos	533	43,55
SVs correctamente etiquetados	**742**	**60,62**
Positivos	462	62,26
Negativos	280	37,74
SVs incorrectamente etiquetados	**133**	**10,87**
Positivos	89	66,92
Negativos	44	33,08
SVs no detectados	**349**	**28,51**
Positivos	140	40,11
Negativos	209	59,89

Tabla 4: Resultados por segmentos

Como resultado obtenemos una precisión de 0,848 y una cobertura de 0,606. Si comparamos estas cifras con las ofrecidas en el apartado anterior, nos encontramos con que no es necesario obtener una exactitud muy alta a nivel de detección de segmentos relevantes para obtener unos resultados de clasificación correcta más que aceptables (89%). Además, nos parece interesante comparar los datos de precisión y cobertura para los casos de segmentos valorativos positivos y negativos por separado, lo que recogemos en la Tabla 5.

Dataset	Precisión	Cobertura
SVs global	0,848	0,616
SVs positivos	0,838	0,669
SVs negativos	0,864	**0,525**

Tabla 5: Datos de precisión y cobertura

Como vemos, existe una diferencia significativa en cuanto a la cobertura de los segmentos valorativos negativos con respecto a los positivos, mientras que la precisión se mantiene en índices parecidos. De esto deducimos que Sentitext tiene más facilidad para detectar los segmentos valorativos positivos que los negativos, lo que nos resulta de mucha utilidad para identificar problemas del analizador. Esta idea queda reforzada por el hecho de que, de los casos en los que existe diferencia entre la valoración del usuario y Sentitext, en aproximadamente el 70% de los mismos Sentitext da una valoración más positiva del texto de la que da el usuario. Típicamente el usuario ha valorado el hotel con una estrella, "pésimo", y Sentitext lo evalúa con dos estrellas, "malo".

El siguiente paso es investigar las causas de esta disparidad y, en general, entender las causas de errores y omisiones.

6 Análisis de errores

Además de cuantificar los resultados, nos ha sido muy útil clasificar los errores y omisiones del analizador con el fin de obtener una idea más precisa de aquellos aspectos concretos que pueden ser mejorables. Para ello hemos llevado a cabo un análisis pormenorizado de los textos, desde el punto de vista formal, semántico y discursivo. De estos tres apartados, la fuente de error (y sobre todo de omisión) más común proviene sin duda del segundo.

6.1 Aspectos formales

Existen una serie de elementos formales que determinan el sentido de la valoración.

- Repeticiones de palabras o letras: "habitación muy muy pequeña", "el ascensor es lento, leennttoo, llleeennntttooo".
- Mayúsculas: de esta misma forma, el uso de la mayúscula en "la habitación era MUY PEQUEÑA" tiene como objetivo enfatizar este mismo aspecto negativo.
- Signos de puntuación: "¡¡¡¡No vayas!!!!", que tratan de representar gráficamente lo que en lengua hablada sería una entonación enfática.
- Errores ortográficos. Ej. "venia" en lugar de "venía". Como hemos indicado, realizamos una corrección previa utilizando un corrector ortográfico, incapaz de detectar errores de este tipo.
- Palabras en otro idioma. Por último, dado que las críticas se refieren a hoteles de Londres, hay usuarios que tienden a usar palabras inglesas en su descripción del hotel. Por ej. "no te ponen amenities".
- Extensión del texto. Paradójicamente, hemos detectado que la valoración global de Sentitext es menos precisa cuando el texto es o muy corto o muy largo. En el primer caso se debe a que no existen muchos elementos léxicos con carga afectiva con los que poder obtener una media representativa del conjunto. En el caso de los textos demasiado largos, nos encontramos en muchas ocasiones con que el usuario divaga y se desvía del objeto principal de su crítica, o bien lo compara con otros hoteles que valora de forma opuesta.

Por otra parte, aunque en los textos de nuestra muestra no se nos ha dado el caso, existen muchos otros dominios en lo que sí es muy habitual la utilización de elementos textuales que expresan opinión o emoción, como son los *emoticonos*. En definitiva, un tratamiento adecuado de estos elementos formales mejoraría, en una proporción variable según el origen del texto, el resultado del análisis.

6.2 Semántica de dominio específico

Como se muestra en la Tabla 4, el porcentaje de segmentos valorativos no detectados por Sentitext es considerablemente mayor que los que sí detecta pero etiqueta erróneamente (con polaridad opuesta), de ahí la significativa diferencia entre las cifras de precisión y cobertura. Fundamentalmente, esto se debe a que los lexicones de los que se sirve la aplicación no han sido diseñados para analizar dominio especializado alguno, por tanto, las entradas léxicas que contienen estos recursos son aplicables en cualquier contexto.

Existen un buen número de contenidos valorativos expresados por los usuarios que lo son únicamente en el contexto de los hoteles y quizás en otros, pero no en todos, por lo que no han sido recogidos en los lexicones de Sentitext. Estos elementos, de los que ofrecemos algunos ejemplos a continuación, son de unos tipos muy concretos.

- **Tener (+) vs. No tener (-)**. Ej.: "hay wi-fi en las habitaciones" o "el cuarto tiene televisión de plasma de 26 pulgadas" serían consideradas como evaluaciones positivas

en este dominio particular por cualquier lector, a diferencia de "no tenía aire acondicionado" o "no hay ascensor".

- **Grande (+) vs. Pequeño (-)**. Ej.: "baño de caravana, es decir ducha, lavabo y retrete todo en un metro cuadrado", "nos alojamos en habitación superior y era enorme".
- **Nuevo (+) vs. Viejo (-)**. Ej.: "los grifos viejísimos", "el aseo de nuestra habitación era de hace más de 20 años", "es nuevo, todo limpio, ropa de cama nueva", "el hotel es moderno y agradable".
- **Cálido (+) vs. Frío (-)**. Ej.: "café bastante malo o infusiones y cereales con la leche fría", "el agua caliente se terminaba", "el agua de la ducha sale todo lo caliente que se quiera y con presión", "las habitaciones están calentitas".
- **Limpieza (+) vs. Suciedad (-)**. Aunque palabras y frases relacionadas con la limpieza y la suciedad están contempladas en nuestros lexicones con la valencia adecuada, aspectos como cambiar o reponer sábanas, toallas y demás productos de aseo suscitan buenas y malas críticas. Ej.: "diariamente cambiaban toallas y reponían geles", "no cambiaban las sábanas en los 3 días que estuvimos".
- **Cercanía (+) vs. Lejanía (-)**. Ej. "muy cerca del metro", "a dos pasos de la Torre de Londres y una boca de metro también muy cerca", "te lo venden como cercano y está a una hora en autobús del centro".

Aunque existen algunos más, lo que nos parece interesante es que la clasificación que acabamos de mostrar podría servir de base para desarrollar un módulo específico del dominio. Aunque en estos momentos no tenemos planeado desarrollarlo, no descartamos hacerlo en el futuro. La relevancia del dominio, como mencionábamos anteriormente, es elevada. Pensamos que los resultados del analizador mejorarían ostensiblemente si contáramos con un tratamiento automatizado adecuado del mismo.

6.3 Aspectos discursivos

El análisis del discurso de los textos nos da las claves para entender por qué Sentitext ofrece una mejor detección de segmentos positivos que negativos.

En concreto, hemos detectado tres fenómenos discursivos que tienden a dificultar la tarea del analizador.

6.3.1 Utilización de lenguaje indirecto.

La utilización del lenguaje indirecto es algo absolutamente generalizado en la lengua, pero sin duda se usa con mayor incidencia para expresar opiniones negativas que positivas, es decir, la positividad tiende a expresarse de una manera más directa y por tanto es más fácil de detectar mediante el uso del vocabulario, que es lo que a fin de cuentas mide Sentitext.

6.3.2 La ironía y el sarcasmo.

Estos mecanismos expresivos no son demasiado recurrentes en las críticas on-line, debido fundamentalmente a que discursivamente es necesario disponer de un contexto amplio para que sean entendidos, y los textos de valoraciones suelen ser de corta extensión. No obstante, nos encontramos con casos como "si alguien se atreve a vivir esta maravillosa experiencia", donde *maravillosa* viene determinado por un contexto de tipo negativo. Igualmente, la palabra *lujo* es usada en el siguiente fragmento con significado contrario al literal: "nos dicen que nos van a poner 2 individuales y la segunda no nos la cobran, uaauauuuuu qué lujo!!!", "nada más entrar en la recepción sientes 'un perfume embriagador' a fritanga vomitivo". Resulta interesante el uso de una oración coordinada disyuntiva en las que se repite la misma oración como recurso ironizante: "había mermelada de blackcurrant o mermelada de blackcurrant".

6.3.3 La *cortesía lingüística*

Hay una tendencia generalizada por parte de los usuarios a ser políticamente correctos. En general, tendemos a suavizar nuestras críticas. El principio de cortesía lingüística (Brown y Levinson, 1987) explicaría en gran medida las diferencias detectadas. Este principio establece que los hablantes tenemos una tendencia natural a mostrar una imagen positiva de nosotros mismos. Igualmente, considera que los oyentes contribuimos a que nuestros interlocutores mantengan esa imagen positiva de ellos mismos.

Este principio explicaría muchos casos de disparidad entre las opiniones vertidas por los usuarios y sus propias valoraciones, que son más de las que cabría esperar.

En ocasiones, sin embargo, se produce el efecto contrario a la cortesía lingüística. El usuario infravalora el hotel. Se dan casos en los que de principio a fin el autor no hace otra cosa

que enumerar cualidades positivas del hotel. Resaltan especialmente fórmulas en la primera y la última frase tales como: "ha sido una buena elección" y "Sin duda un sitio para volver". Resulta incomprensible asignar 4 estrellas en lugar de 5 a aquellos textos que las usan. Desde el punto de vista del discurso periodístico, la polaridad del último párrafo coincide a grandes rasgos con la polaridad global del texto. Empezar y acabar un texto con opiniones positivas o negativas realza la polaridad positiva o negativa respectivamente.

Desde el punto de vista pragmático, y desde una perspectiva global, se observa un mecanismo de amortiguación, especialmente de los textos valorados con 2 y 3 estrellas. Los usuarios tienden a suavizar con recursos pragmáticos y retóricos las opiniones muy negativas. Este hecho a menudo contrasta con la valoración final. Quizá el principio de cortesía lingüística pueda, al menos parcialmente, responder a este fenómeno: los seres humanos tratamos de proyectar una imagen propia positiva. De ahí que, por ejemplo, modifiquemos el tenor de nuestro registro dependiendo de nuestros interlocutores y de los contextos en los que se desarrolla la comunicación.

Desde el punto de vista del discurso periodístico, la polaridad del último párrafo coincide a grandes rasgos con la polaridad global del texto. Empezar y acabar un texto con opiniones positivas o negativas realza la polaridad positiva o negativa respectivamente.

7 Conclusiones

A la vista de los resultados obtenidos, parece evidente que el rendimiento del software es más que aceptable. Incluso con las limitaciones discutidas, la tasa de acierto en cuanto a la polaridad es muy alta. Concluimos que, en lo que se refiere al contexto discursivo de la crítica de productos y servicios, el léxico es sin duda el elemento discursivo que en mayor medida codifica el eje axiológico. Por tanto, es imprescindible disponer de fuentes de conocimiento léxico de gran cobertura y alta calidad.

No obstante, queda aún mucho por hacer para lograr un índice de acierto cercano al 100%. Pensamos que para lograr este objetivo necesitamos avanzar en dos direcciones. Primero, desarrollar las reglas de contexto y optimizar su aplicación. En segundo lugar, nos planteamos la posibilidad de crear módulos léxicos específicos de dominios, que se aplicarían únicamente a nivel de léxico como paso previo. También sería interesante implementar maneras de controlar los elementos formales que en muchos textos actúan como marcadores de afectividad.

Otra de las vías que estamos explorando es la mejora de la calificación global del texto mediante un sistema de "prominencia afectiva" que otorga mayor peso a los indicadores afectivos según su posición en el texto, algo que ya ha sido probado por otros autores (Taboada y Grieve, 2004, Taboada et al., 2006).

Finalmente, aún vemos muy lejano el tratamiento automatizado de recursos lingüísticos como las figuras retóricas y otros que implican conocimiento enciclopédico, pero sin duda es necesario estudiar hasta qué punto dichos elementos interfieren en el análisis.

En cualquier caso, lo que nos ha quedado más patente es que, incluso unos niveles de precisión y sobre todo de cobertura relativamente moderados en cuanto a los segmentos valorativos, es suficiente para obtener una clasificación de la polaridad de los textos de opinión cercanos al 90%. Por supuesto, es necesario poner a prueba el software en otros dominios y contextos para comprobar si su rendimiento es similar.

Bibliografía

Atserias, J. et al. 2006. FreeLing 1.3: Syntactic and semantic services in an open-source NLP library. En Proceed*ings of the fifth international conference on Language Resources and Evaluation*. LREC 2006. Genoa, Italy: ELRA.

Aue, A. & Gamon, M. 2005. Customizing Sentiment Classifiers to New Domains: A Case Study. In Recent Advances in Natural Language Processing (RANLP). Borovets, Bulgaria.

Boldrini, E. et al. 2009. EmotiBlog: an Annotation Scheme for Emotion Detection and Analysis in Non-traditional Textual Genres. En *Proceedings of The 2009 International Conference on Data Mining (DMIN 2009)*, Las Vegas, USA: CSREA Press, pp. 491-497.

Brown, P. and Levinson, S. C.. 1987. *Politeness: Some universals in language usage*. Cambridge: Cambridge University

Press.

Cruz, F. et al., Clasificación de documentos basada en la opinión: experimentos con un corpus de críticas de cine en español. *Procesamiento del Lenguaje Natural*, (41), 73-80.

Das, S.R. & Chen, M. 2001. Yahoo! for Amazon: Sentiment Parsing from Small Talk on the Web. En Proceed*ings of the 8th Asia Pacific Finance Association Annual Conference*.

Dave, K., Lawrence, S. & Pennock, D.M. 2003. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. En Proceed*ings of the 12th international conference on World Wide Web*. Budapest, Hungary: ACM, pp. 519-528.

Esuli, A. & Sebastiani, F. 2006. SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. En Proceed*ings of the 5th Conference on Language Resources and Evaluation (LREC '06)*. LREC 2006. pp. 417-422.

Gamon, M. 2004. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. En Proceed*ings of the 20th international conference on Computational Linguistics*. Geneva, Switzerland: Association for Computational Linguistics, p. 841.

Liu, H., Selker, T. & Lieberman, H. 2003. Visualizing the affective structure of a text document. In *CHI '03 extended abstracts on Human factors in computing systems*. Ft. Lauderdale, Florida, USA: ACM, pp. 740-741.

Pang, B. & Lee, L. 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2), 1-135.

Pang, B. & Lee, L. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. En Proceed*ings of ACL 2005*. ACL. pp. 115-124.

Pitel, G. & Grefenstette, G. 2008. Semi-automatic Building Method for a Multidimensional Affect Dictionary for a New Language. En Proceed*ings of the Sixth International Language Resources and Evaluation*. LREC '08. Marrakech, Morocco: ELRA.

Polanyi, L. & Zaenen, A. 2006. Contextual Valence Shifters. In *Computing Attitude and Affect in Text: Theory and Applications*. Dordrecht, The Netherlands: Springer, pp. 1-10.

Taboada, M., Anthony, C. & Voll, K. 2006. Methods for Creating Semantic Orientation Dictionaries. En Proceed*ings of 5th International Conference on Language Resources and Evaluation (LREC)*. Genoa, Italy: ELRA.

Taboada, M. & Grieve, J. 2004. Analyzing Appraisal Automatically. In *AAAI Technical Report SS-04-07*. American Association for Artificial Intelligence Spring Symposium on Exploring Attitude and Affect in Text. Stanford, pp. 158-161.

Tan, S. et al. 2007. A novel scheme for domain-transfer problem in the context of sentiment analysis. En Proceed*ings of the sixteenth ACM conference on Conference on information and knowledge management*. Lisbon, Portugal: ACM, pp. 979-982.

Thet, T.T. et al. 2009. Sentiment analysis of movie reviews on discussion boards using a linguistic approach. En Proceed*ing of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*. Hong Kong, China: ACM, pp. 81-84.

Thomas, M., Pang, B. & Lee, L. 2006. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. En Proceed*ings of EMNLP*. EMNLP. pp. 327-335.

Turney, P.D. & Littman, M.L. 2002. Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus.

Turney, P.D. 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. En Proceed*ings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*. ACL 2002. Philadelphia, USA., pp. 417-424.

Wilson, T., Wiebe, J. & Hoffmann, P. 2009. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3), 399-433.