Procesamiento del Lenguaje Natural, Revista nº 50 marzo de 2013, pp 13-20 recibido 23-10-12 revisado 14-01-13 aceptado 19-02-13

Clasificación de polaridad en textos con opiniones en español mediante análisis sintáctico de dependencias

Polarity classification of opinionated Spanish texts using dependency parsing

David Vilares, Miguel A. Alonso y **Carlos Gómez-Rodríguez**
Departamento de Computación, Universidade da Coruña
Campus de Elviña, 15011 A Coruña
{david.vilares, miguel.alonso, carlos.gomez}@udc.es

Resumen: En este artículo se describe un sistema de minería de opiniones que clasifica la polaridad de textos en español. Se propone una aproximación basada en PLN que conlleva realizar una segmentación, tokenización y etiquetación de los textos para a continuación obtener la estructura sintáctica de las oraciones mediante algoritmos de análisis de dependencias. La estructura sintáctica se emplea entonces para tratar tres de las construcciones lingüísticas más significativas en el ámbito que nos ocupa: la intensificación, las oraciones subordinadas adversativas y la negación. Los resultados experimentales muestran una mejora del rendimiento con respecto a los sistemas puramente léxicos y refuerzan la idea de que el análisis sintáctico es necesario para lograr un análisis del sentimiento robusto y fiable.
Palabras clave: Minería de opiniones, Análisis del sentimiento, Análisis sintáctico de dependencias

Abstract: This article describes an opinion mining system that classifies the polarity of Spanish texts. We propose a NLP-based approach which performs segmentation, tokenization and POS tagging of texts to then obtain the syntactic structure of sentences by means of a dependency parser. The syntactic structure is then used to address three of the most significant linguistic constructions in the area in question: intensification, adversative subordinate clauses and negation. Experimental results show an improvement in performance with respect to purely lexical approaches and reinforce the idea that parsing is required to achieve a robust and reliable sentiment analysis system.
Keywords: Opinion Mining, Sentiment Analysis, Dependency Parsing

1. Introducción

El auge en los últimos años de los blogs, los foros y las redes sociales ha hecho que millones de usuarios utilicen estos recursos para expresar sus opiniones sobre toda una variedad de temas. La diversidad y cantidad de críticas presentes en la web resultan de gran utilidad a fabricantes y vendedores, que ven en ellas un mecanismo para conocer de primera mano cómo sus artículos son percibidos por los consumidores. Los beneficios asociados a conocer toda esta información, sumados a la complejidad técnica del análisis de las opiniones, han provocado que se hayan comenzado a demandar soluciones capaces de monitorizar este flujo ingente de reseñas.

Todo ello ha contribuido a que la minería de opiniones (MO), también conocida como análisis del sentimiento, esté jugando un papel importante como ámbito de investigación en los últimos años. La MO se centra en tratar automáticamente información con opinión, lo que permite, entre otras cosas, extraer la polaridad (positiva, negativa, neutra o mixta) de un texto (Pang y Lee, 2008).

En este artículo presentamos un sistema de clasificación de polaridad para textos escritos en español, cuyas principales características son la utilización de diccionarios semánticos y de la estructura sintáctica de las oraciones para clasificar un texto subjetivo como positivo o negativo. La utilidad práctica de esta aproximación viene avalada por los resultados experimentales presentados, que muestran una mejora en precisión de más de cuatro puntos porcentuales con respecto a un sistema reciente que no hace uso de la sintaxis.

ISSN 1135-5948

El resto del artículo se organiza como sigue. En la sección 2 se revisa brevemente la situación actual de la MO, centrándose en lo referido a la detección de la polaridad. En la sección 3 se describe la propuesta planteada y se detallan los aspectos sintácticos tratados. En la sección 4 se muestran detalles de implementación y los resultados de los experimentos realizados. Por último, en la sección 5 se presentan las conclusiones y las principales líneas de trabajo futuras.

2. Estado del arte

Una parte importante de los esfuerzos actuales relacionados con la MO se están realizando en tareas relativas a la clasificación de la polaridad, problema que ha sido abordado desde dos enfoques principales. El primero asume esta tarea como un proceso genérico de clasificación (Pang, Lee, y Vaithyanathan, 2002): a partir de un conjunto de entrenamiento, donde los textos son anotados con su polaridad, se construye un clasificador mediante aprendizaje automático (AA). El segundo enfoque se apoya en la orientación semántica (OS) de las palabras, donde cada término que expresa opinión es anotado con un valor que representa su polaridad (Turney, 2002). Este segundo enfoque es el que tomaremos como base para el desarrollo de nuestro trabajo.

La mayor parte de los sistemas de MO se centran en el tratamiento de textos en inglés. En el caso de textos escritos en español, probablemente el sistema más relevante sea *The Spanish SO Calculator* (Brooke, Tofiloski, y Taboada, 2009), desarrollado en la Universidad Simon Fraser de Canadá. Este sistema, además de resolver la OS almacenada a nivel individual en adjetivos, sustantivos, verbos y adverbios; trata modificadores de la polaridad como son la negación o los intensificadores (*"muy", "poco", "bastante", ...*). También detecta y descarta el sentimiento reflejado en el contenido no fáctico del texto, representado, por ejemplo, mediante expresiones condicionales o subjuntivas.

La manera más habitual de tratar todas estas construcciones lingüísticas es a nivel léxico y en este aspecto The Spanish SO Calculator no es una excepción. En lo que respecta al tratamiento de la negación, (Taboada et al., 2011) utiliza información morfológica para identificar el alcance de la negación, mientras que (Yang, 2008) considera dicho alcance como los términos situados a la derecha de la negación y en (Fernández Anta et al., 2012) se emplea una heurística que asume que los tres elementos a continuación de una negación son los que deben cambiar su polaridad. Para la intensificación, (Fernández Anta et al., 2012) considera de nuevo que los tres términos a la derecha son los que deben variar su polaridad. (Taboada et al., 2011) además de los intensificadores propiamente dichos, trata como tales aspectos del discurso como la conjunción *"pero"* o las mayúsculas.

Nuestra propuesta sigue una estrategia distinta, que se basa en obtener la estructura sintáctica del texto para tratar las construcciones lingüísticas e identificar los elementos de la frase que están implicados en ellas. A este respecto, trabajos anteriores (Jia, Yu, y Meng, 2009) ya han mostrado los beneficios de utilizar la estructura sintáctica de la frase en aquellos textos en los que aparecen ocurrencias de términos negativos.

Un problema adicional al que se enfrentan los sistemas de MO es la calidad ortográfica de los textos a analizar. Cuando éstos provienen de la web, debe tenerse en cuenta que es frecuente que sus autores omitan acentos, letras o vocablos; o empleen tanto abreviaturas no reconocidas como oraciones agramaticales. La solución más utilizada consiste en emplear patrones heurísticos para adaptar el texto (Saralegi Urizar y San Vicente Roncal, 2012; Martínez Cámara et al., 2012) .

3. Clasificación de opiniones basada en dependencias sintácticas

En contraste con las propuestas léxicas dominantes hasta el momento, en este trabajo proponemos la utilización de la estructura sintáctica de la frase para obtener la OS de un texto. Como primer paso, es necesario preprocesar los textos, para ello se ha diseñado un preprocesador *ad-hoc* que trata los siguientes aspectos:

- La unificación de expresiones compuestas, que actúan como una sola unidad de significado (*"a menos que", "en absoluto",...*).
- La normalización de los signos de puntuación. En un entorno web es común obviar las normas ortográficas respecto a la colocación de signos como el punto o la coma, lo que puede afectar negativamente al resto del procesado.

A continuación, se procede a segmentar el texto en oraciones y a dividir cada una de ellas en *tokens* (principalmente para palabras, pero también signos de puntuación, números, etc.) para después realizar la etiquetación morfosintáctica de cada una de las palabras del texto.

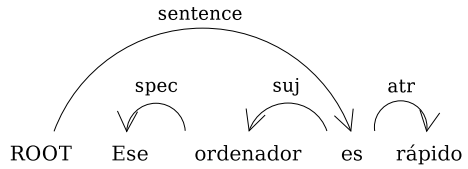

Figura 1: Ejemplo de árbol de dependencias

El siguiente paso consiste en realizar el análisis sintáctico de dependencias mediante el cual se identifican relaciones binarias padre/dependiente entre los términos de una oración. Se incluye un enlace con un elemento artificial inicial (ROOT) para facilitar las definiciones formales e implementaciones. Cada uno de esos vínculos binarios constituye una dependencia, que se anota con la función sintáctica que relaciona los dos términos. A la estructura obtenida se le denomina árbol de dependencias. En la Figura 1 se ilustra un ejemplo sencillo de este tipo de análisis. Como corpus de referencia para la definición de las relaciones de dependencia se ha utilizado Ancora (Taulé, Martí, y Recasens, 2008).

Finalmente, para la realización del análisis semántico, nuestra propuesta se apoya en el SODictionariesV1.11Spa (Brooke, Tofiloski, y Taboada, 2009). Se trata un conjunto de diccionarios de polaridad para adjetivos, sustantivos, verbos, adverbios e intensificadores; cuyo contenido se resume en la Tabla 1. Cada término se encuentra anotado con un valor entre -5 y 5, donde -5 es lo más negativo y 5 lo más positivo. El valor asignado a cada palabra se corresponde con una orientación semántica genérica, independientemente del dominio o contexto en el que se utilice. Así, por ejemplo, al adjetivo *"rápido"* o al verbo *"recomendar"* se les asocia una polaridad de valor 2. Es importante señalar que los valores numéricos asociados a los intensificadores tienen un significado distinto, ya que representan el porcentaje (positivo o negativo) por el que modifican el sentimiento de la expresión a la que afectan.

Diccionario	**Nº términos**
adjetivos	2,049
sustantivos	1,324
verbos	739
adverbios	548
intensificadores	157

Tabla 1: Contenido del SODictionariesV1.11Spa

3.1. Propuesta base

Nuestra versión inicial determina la polaridad de un texto únicamente a partir de la combinación de la OS de sustantivos, adjetivos, verbos y adverbios; esto es, sin considerar ninguna construcción lingüística compleja, lo que equivale a ignorar la estructura sintáctica del texto. En la Figura 2 se ilustra un ejemplo de análisis de la OS sobre el árbol de dependencias correspondiente a la oración *"Ese ordenador es muy rápido, pero no recomiendo comprarlo"*. Podemos observar que la propuesta base establece una OS muy positiva para un texto que intuitivamente se percibe como ligeramente negativo. Se trata de un ejemplo didáctico que refleja los problemas de obviar fenómenos como la intensificación, los nexos adversativos o la negación a la hora de extraer completamente la polaridad.

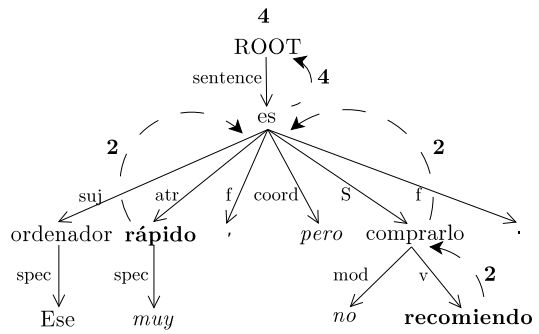

Figura 2: Análisis semántico sobre árbol de dependencias

3.2. Tratamiento de la intensificación

Los intensificadores son términos o expresiones que modifican la polaridad de ciertas palabras. Consideraremos dos tipos: *amplificadores*, si permiten aumentar la polaridad (*"muy", "bastante",...*), y *decrementadores* si la disminuyen (*"poco", "en absoluto",...*). Para modelar esta construcción se asocia a cada

intensificador un factor de ponderación. Así, basándonos en el SODictionariesV1.11Spa, al amplificador *"muy"* se le asocia el valor 0,25 y al decrementador *"en absoluto"*, -1. La principal diferencia radica en que nuestra propuesta utiliza el árbol de dependencias para determinar la parte de la frase que se ve afectada por tal modificación, considerando las dependencias anotadas en Ancora como *spec, espec, cc* o *sadv*.

Para el ejemplo presentado en la Figura 2, la OS de *"muy rápido"* se obtendría incrementando en un 25 % la OS de *"rápido"*: $2 * (1 + 0,25) = 2,5$. En caso de que haya varios intensificadores, se combinan todos sus porcentajes de intensificación antes de que actúen sobre el término afectado. Por ejemplo, si la expresión intensificada fuese *"en absoluto muy rápido"* la OS se obtendría como $2 * (1 + (-1 + 0,25)) = 0,5$.

En un entorno web existen otras formas de enfatizar opiniones, como son el empleo de mayúsculas o de exclamaciones. Hemos tratado estas peculiaridades siguiendo un enfoque similar al del resto de intensificadores.

3.3. Tratamiento de las oraciones adversativas

Los nexos adversativos permiten contraponer hechos expresados en dos oraciones. En un entorno de MO este tipo de frases se emplean para restringir o excluir opiniones, lo que puede ser considerado como un caso especial de intensificación. Disponer de un árbol de dependencias resulta de gran utilidad en este caso, ya que nos permite identificar con precisión tanto la oración subordinada como la subordinante. Desafortunadamente, el corpus de Ancora representa sintácticamente este tipo de oraciones de forma distinta según el nexo concreto utilizado, por lo que el tratamiento realizado para este tipo de cláusulas no ha sido todo lo completo que nos hubiera gustado. Hemos optado por centrarnos en los nexos más relevantes que Ancora representa de manera uniforme. Se han dividido estos nexos en dos grupos: los *restrictivos*, que reducen la OS de la oración principal y donde destaca la conjunción *"pero"*; y los *excluyentes*, que eliminan enteramente lo expresado en la primera oración, entre los que se encuadran conjunciones como *"sino"*. Así, según la clase de nexo, se pondera el sentimiento acumulado, tanto en la oración subordinante como en la subordinada, de forma distinta. En la Tabla 2 se ilustran los factores de ponderación $F_{principal}$ y $F_{subordinada}$, establecidos mediante una evaluación empírica del SFU Spanish Review Corpus, cuyo contenido se detalla en la sección 4.2.

Nexo	$F_{principal}$	$F_{subordinada}$
Restrictivo	0,75	1,4
Excluyente	0	1

Tabla 2: Factores de ponderación según el tipo de nexo adversativo

Para homogeneizar en un futuro la estructura sintáctica de otras subordinadas adversativas, y para simplificar la ponderación de estas oraciones; se optó por reestructurarlas en el árbol de dependencias. En la Figura 3 se ilustra la estructura esquemática de una oración adversativa una vez reorganizada. Se observa que en el nivel superior de la cláusula subordinada se incluye un nodo de apoyo, representado por **. Se crea también un nuevo tipo de de dependencia, *art_rel_adversative*, para identificar sintácticamente el inicio de una cláusula de este tipo. Si se retoma el ejemplo de la Figura 2, donde aparecen dos oraciones conectadas por la conjunción adversativa *"pero"*, la estructura sintáctica reorganizada sería la ilustrada en la Figura 4.

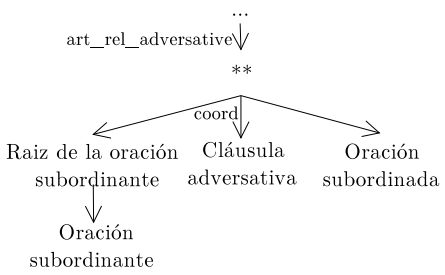

Figura 3: Reestructuración de oraciones adversativas

3.4. Tratamiento de la negación

Son muchos los términos o expresiones que permiten negar una opinión. Sin embargo, la frontera entre un negador como tal y un intensificador decrementador es difusa. En este trabajo se ha restringido el tratamiento de este fenómeno a los términos *"no"*, *"nunca"* y *"sin"*. Otras expresiones negadoras, como *"lo menos"* o *"en absoluto"*, han sido tratadas como intensificadores. Para ello, se ha aprovechado la información semántica proporcio-

nada por el SODictionariesV1.11Spa para este tipo de locuciones.

Para resolver el sentimiento de una oración con ocurrencias de términos negativos es necesario realizar dos pasos: identificar el alcance de la negación y modificar la polaridad del fragmento de la oración correspondiente.

3.4.1. Identificación del alcance de la negación

Nuestra estrategia para identificar el alcance de la negación se basa en la propuesta de (Jia, Yu, y Meng, 2009). Sin embargo, el procedimiento ha sido adaptado a las peculiaridades del análisis sintáctico realizado. Las características del árbol de dependencias permiten definir un procedimiento estrictamente sintáctico, basado en las relaciones entre elementos, sin necesidad de localizar delimitadores léxicos.

La forma de identificar ese alcance difiere según el negador utilizado. Cuando se emplea el término *"sin"*, el árbol de dependencias nos asegura que la rama descendiente constituye el alcance de ese negador, sin necesidad de analizar el tipo de relación. Por contra, la estructura sintáctica utilizada para representar los elementos *"no"* y *"nunca"*, requiere identificar dependencias concretas como *neg* o *mod*, e iniciar un proceso más complejo. En primer lugar, se establece un alcance candidato, formado tanto por el padre del negador como por sus hermanos. A continuación se corrige dicho alcance aplicando una serie de reglas heurísticas, que son procesadas en orden hasta que una cumpla los requisitos:

1. *Regla del padre subjetivo*: Si el padre del negador aparece en los diccionarios semánticos, entonces sólo él constituye el alcance corregido de la negación.

2. *Regla del atributo o complemento directo*: Si alguno de los hermanos desempeña una de estas funciones sintácticas, entonces dicho hermano forma parte del alcance de la negación.

3. *Regla del complemento circunstancial más cercano*: Si alguna rama al mismo nivel del negador actúa cómo complemento circunstancial, entonces dicha rama forma el alcance corregido. En caso de varios complementos circunstanciales, sólo se incorpora el más cercano físicamente al negador.

Si ninguna regla se cumple, entonces se asume el alcance candidato (salvo el nodo padre) como el corregido. En el ejemplo de la Figura 4, para la negación *"no recomiendo comprarlo"*, ninguna de la reglas se cumple, por lo que el alcance corregido estaría formado sólo por el verbo *"recomiendo"*.

3.4.2. Modificación de la polaridad

Nuestra propuesta para resolver la modificación de la polaridad que implica una negación es similar a la empleada en trabajos como (Taboada et al., 2011). Una vez obtenido el alcance corregido, se extrae su polaridad, y a continuación, el valor obtenido es modificado en una cantidad preestablecida de signo contrario. Para los negadores *"no"* y *"nunca"*, dicho valor es 4, mientras que para *"sin"* el valor es menor, 3,5, para ajustarse a su carácter más local. Así, en el ejemplo de la Figura 4, se observa como para la negación de *"recomiendo"* se obtiene una OS de -2.

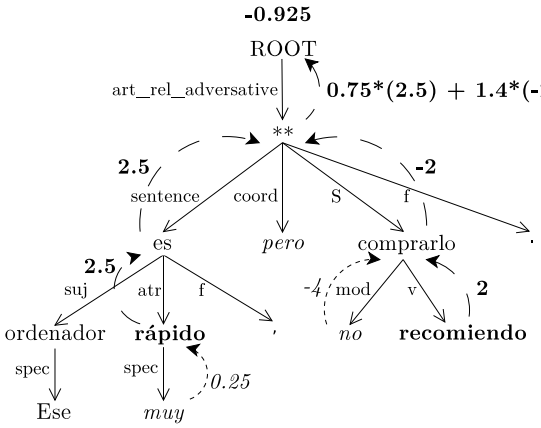

Figura 4: Análisis final de la OS sobre el árbol de dependencias reestructurado

4. Resultados experimentales

4.1. Implementación

Nuestra propuesta para la clasificación de la polaridad se ha implementado en Python, apoyado en el *toolkit* NLTK[1] para las tareas de segmentación, tokenización y etiquetación. En concreto, para la tarea de etiquetación se ha aplicado el algoritmo de Brill utilizando el corpus Ancora (Taulé, Martí, y Recasens, 2008) para el entrenamiento (se ha utilizado el 90 % del corpus para el entrenamiento y el 10 % restante para la evaluación). Para mejorar el rendimiento práctico del etiquetador

[1] http://nltk.org/

sobre el análisis de textos de la web, donde se obvian los acentos en muchas palabras, el fragmento del corpus destinado al aprendizaje fue ampliado de forma que cada oración dispusiese de su equivalente sin palabras acentuadas gráficamente. Los resultados de la evaluación del etiquetador, mostrados en la Tabla 3, sugieren que las ambigüedades creadas por esta duplicación apenas afectan a la precisión teórica del etiquetador y, sin embargo, se comprobó empíricamente que mejoraba la anotación sobre textos no acentuados.

Corpus	Precisión
Original	0,9586
Ampliado	0,9571

Tabla 3: Precisión del etiquetador de Brill

La tarea del análisis sintáctico de dependencias se ha realizado con el algoritmo *Nivre arc-eager* (Nivre, 2008) generado con MaltParser[2] (Nivre et al., 2007) mediante aprendizaje automático a partir del corpus Ancora.

En la sección anterior se comentó cómo se han tratado algunas construcciones de naturaleza sintáctica, sin embargo, hay aspectos que no pueden resolverse a ese nivel. Ejemplo de ello es la mayor importancia de las oraciones finales de una opinión. Para modelar esta peculiaridad, en nuestra propuesta se ha optado por aumentar en un 75 % la OS de las tres últimas frases de una crítica.

Otro aspecto a tener en cuenta es el introducido en (Kennedy y Inkpen, 2006), donde se habla del problema de la tendencia positiva del lenguaje humano. 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"* son dos ejemplos de esta situación. Para compensar dicha desviación, muchas aproximaciones léxicas incrementan la OS de los términos negativos, mejorando notablemente su rendimiento. Sin embargo, el empleo de esta técnica en nuestra propuesta resultó contraproducente. Sí se consiguió mejorar la precisión del sistema aumentando la dispersión de las OS de sustantivos, adjetivos, verbos y adverbios del SODictionariesV1.11Spa en un 20 %, esto es, que sus polaridades comprendan valores entre -6 y 6. Todos los aspectos que incrementaron el rendimiento se incluyeron en la versión final de nuestro sistema.

4.2. Evaluación

Para la evaluación de nuestra propuesta se ha empleado un corpus formado por 400 documentos: el SFU Spanish Review Corpus (Brooke, Tofiloski, y Taboada, 2009). Contiene reseñas de productos y servicios de ocho categorías distintas: lavadoras, hoteles, películas, coches, ordenadores, libros, música y móviles. Para cada categoría se dispone de un total de 50 documentos, donde en 25 de ellos se expresa una opinión positiva, mientras los otros 25 expresan una negativa.

Nuestra propuesta procesa cada texto y obtiene como resultado su OS, si ésta es mayor que 0 el texto se cataloga como positivo, en caso contrario como negativo. En la Tabla 4 se ilustra la precisión para distintas configuraciones. Todas las construcciones lingüísticas tratadas han mejorado el rendimiento. Especialmente significativo es el incremento obtenido con la incorporación de la negación. Se realizaron test chi-cuadrado ($p < 0,01$), comparando con las polaridades correctas. Con un * se ilustran las configuraciones para las que se obtuvieron polaridades que no difieren de manera estadísticamente significativa de las correctas.

Propuesta	Precisión
Base	0,618
+ intensificación	0,660
+ adversativas	0,670
+ negación	0,755*
Final	0,785*

Tabla 4: Precisión al incorporar distintas funcionalidades

Haber utilizado para la evaluación el mismo corpus y los mismos diccionarios semánticos que la solución léxica The Spanish SO-Calculator, permite comparar nuestra alternativa sintáctica con ella. En la Tabla 5 se contrasta el rendimiento. Nuestra propuesta incrementa en un 5,72 % el rendimiento obtenido por The Spanish SO-CAL. También se construyó un clasificador SVM, basado en AA, empleando para ello WEKA[3]. Para su desarrollo, se utilizó el SFU Spanish Review Corpus y como método de evaluación se optó por

[2] http://www.maltparser.org/

[3] http://www.cs.waikato.ac.nz/ml/weka/index.html

una validación cruzada de 10 iteraciones. Todos los términos se cambiaron a su forma minúscula y se utilizó su frecuencia absoluta de aparición. (Brooke, Tofiloski, y Taboada, 2009) también propone un sistema de AA, incluyendo PLN, pero sus resultados no mejoran los presentados con nuestra configuración.

Método	Precisión (%)
Nuestra propuesta	**78,50**
The Spanish SO-CAL	74,25
SVM	72,50

Tabla 5: Precisión para distintos métodos

En la Tabla 6 se muestra la precisión de la versión final del sistema, desglosada para las categorías del corpus. Para los ámbitos considerados de entretenimiento, como las películas o los libros; el rendimiento es peor que la media. Hay dos razones posibles. La primera es referida al empleo de OS genéricas. Términos como *"guerra"* o *"asesino"* son manifiestamente negativos, sin embargo, en dominios relacionados con las novelas o las películas, probablemente describan la temática o el argumento, sin afectar a la calidad del producto. El segundo motivo está relacionado con los gustos personales, lo que complica clasificar la polaridad de ciertos términos en estos ámbitos. Por el contrario, se obtienen mejores resultados en dominios donde los criterios de calidad están claramente establecidos, como es el caso de los hoteles o los ordenadores.

Categoría	Neg	Pos	Total
Lavadoras	0,79	0,86	0,82
Hoteles	0,88	0,92	0,90
Películas	0,67	0,65	0,66
Coches	0,77	0,71	0,74
Ordenadores	0,91	0,82	0,86
Libros	0,80	0,70	0,74
Música	0,84	0,71	0,76
Móviles	0,86	0,76	0,80

Tabla 6: Precisión según categoría

El sistema, con la misma configuración, se evaluó también sobre HOpinion[4] (críticas de hoteles) y sobre CorpusCine (Cruz, Troyano, y Ortega, 2008), para los que se obtuvo una precisión global de 0,89 y 0,64, respectivamente. Es interesante reseñar que estos resultados son similares a los obtenidos para las categorías de hoteles y películas, respectivamente, en el SFU Spanish Review.

5. Conclusiones y trabajo futuro

Este artículo describe una estrategia para resolver la OS de textos con opinión empleando técnicas de análisis de dependencias. Los experimentos realizados confirman que la utilización de la sintaxis resulta útil a la hora de tratar construcciones lingüísticas en un entorno de MO, como son la negación, la intensificación y las frases adversativas. A este respecto, el análisis que se ha hecho de la negación evita contrarrestar artificialmente la tendencia positiva del lenguaje humano. Esto nos sugiere que se está realizando una identificación fiable del alcance de la negación.

En busca de futuras mejoras, tratar las expresiones y construcciones desiderativas es una línea de trabajo que nos gustaría explorar. También se ha planeado realizar una evaluación más exhaustiva con otros algoritmos de análisis sintáctico de dependencias, como el 2-planar (Gómez-Rodríguez y Nivre, 2010).

La evaluación de nuestra propuesta se realizó sobre un corpus de textos extensos creado por (Brooke, Tofiloski, y Taboada, 2009). Al respecto, el éxito de redes como Twitter ha aumentado el interés por analizar textos cortos (Villena-Román et al., 2013), por lo que sería interesante poder evaluar y adaptar nuestro sistema a las características de este tipo de documentos.

Ciertos factores que afectan a la clasificación de la polaridad no se han considerado. Por ejemplo, el problema de la polaridad cambiante para determinados términos según el dominio en el que aparezcan (Pang y Lee, 2008). La ironía o el sarcasmo son dos figuras literarias que se utilizan para expresar una opinión de una forma mucho más creativa y sutil, lo que dificulta su tratamiento y su identificación. A este respecto, en (Reyes y Rosso, 2011) se describe una aproximación para detectar la ironía que podría ser utilizada para enriquecer nuestra propuesta.

Agradecimientos

Este trabajo ha sido parcialmente financiado por el Ministerio de Economía y Competitividad y FEDER (TIN2010-18552-C03-02) y por la Xunta de Galicia (CN2012/008,

[4]http://clic.ub.edu/corpus/hopinion

CN 2012/319).

Bibliografía

Brooke, J., M. Tofiloski, y M. Taboada. 2009. Cross-Linguistic Sentiment Analysis: From English to Spanish. En *Proceedings of the International Conference RANLP-2009*, páginas 50–54, Borovets, Bulgaria. ACL.

Cruz, F., J. A. Troyano, y J. Ortega. 2008. Clasificación de documentos basada en la opinión: experimentos con un corpus de críticas de cine en español. En *Procesamiento de lenguaje natural, 41*, páginas 81–87.

Fernández Anta, A., P. Morere, L. Núñez Chiroque, y A. Santos. 2012. Techniques for Sentiment Analysis and Topic Detection of Spanish Tweets: Preliminary Report. En *TASS 2012 Working Notes*, Castellón, Spain.

Gómez-Rodríguez, C. y J. Nivre. 2010. A transition-based parser for 2-planar dependency structures. En *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL'10, páginas 1492–1501, Stroudsburg, PA, USA. ACL.

Jia, L., C. Yu, y W. Meng. 2009. The effect of negation on sentiment analysis and retrieval effectiveness. En *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM'09, páginas 1827–1830, New York, NY, USA. ACM.

Kennedy, A. y D. Inkpen. 2006. Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, 22(2):110–125.

Martínez Cámara, E., M. T. Martín Valdivia, M. A. García Cumbreras, y L. A. Ureña López. 2012. SINAI at TASS 2012. En *TASS 2012 Working Notes*, Castellón, Spain.

Nivre, J. 2008. Algorithms for deterministic incremental dependency parsing. *Compuational Linguistics*, 34(4):513–553.

Nivre, J., J. Hall, J. Nilsson, A. Chanev, G. Eryigit, S. Kübler, S. Marinov, y E. Marsi. 2007. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.

Pang, B. y L. Lee. 2008. *Opinion Mining and Sentiment Analysis*. now Publishers Inc., Hanover, MA, USA.

Pang, B., L. Lee, y S. Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. En *Proceedings of EMNLP*, páginas 79–86.

Reyes, A. y P. Rosso. 2011. Mining subjective knowledge from customer reviews: a specific case of irony detection. En *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '11, páginas 118–124, Stroudsburg, PA, USA. ACL.

Saralegi Urizar, X. y I. San Vicente Roncal. 2012. Detecting Sentiments in Spanish Tweets. En *TASS 2012 Working Notes*, Castellón, Spain.

Taboada, M., J. Brooke, M. Tofiloski, K. Voll, y M. Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Taulé, M., M. A. Martí, y M. Recasens. 2008. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. En Nicoletta Calzolari Khalid Choukri Bente Maegaard Joseph Mariani Jan Odjik Stelios Piperidis, y Daniel Tapias, editores, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Turney, P. D. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. En *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, páginas 417–424, Stroudsburg, PA, USA. ACL.

Villena-Román, J., S. Lana-Serrano, J.C. González Cristóbal, y E. Martínez-Cámara. 2013. TASS Worshop on Sentiment Analysis at SEPLN. *Procesamiento de Lenguaje Natural, 50.*

Yang, K.. 2008. WIDIT in TREC 2008 blog track: Leveraging multiple sources of opinion evidence. En E. M. Voorhees y Lori P. Buckland, editores, *NIST Special Publication 500-277: The Seventeenth Text REtrieval Conference Proceedings (TREC 2008).*