Procesamiento del Lenguaje Natural, Revista nº 54, marzo de 2015, pp 21-28 recibido 24-10-14 revisado 27-01-15 aceptado 12-02-15

eSOLHotel: Generación de un lexicón de opinión en español adaptado al dominio turístico

eSOLHotel: Building an Spanish opinion lexicon adapted to the tourism domain

M. Dolores Molina González, Eugenio Martínez Cámara, M. Teresa Martín Valdivia, Salud M. Jiménez Zafra
Departamento de Informática, Escuela Politécnica Superior de Jaén
Universidad de Jaén, E-23071 - Jaén
{mdmolina, emcamara, maite, sjzafra}@ujaen.es

Resumen: Desde que la web 2.0 es el mayor contenedor de opiniones en todos los idiomas sobre distintos temas o asuntos, el estudio del Análisis de Sentimientos ha crecido exponencialmente. En este trabajo nos centramos en la clasificación de polaridad de opiniones en español y se presenta un nuevo recurso léxico adaptado al dominio turístico (eSOLHotel). Este nuevo lexicón usa el enfoque basado en corpus. Se han realizado varios experimentos usando una aproximación no supervisada para la clasificación de polaridad de las opiniones en la categoría de hoteles del corpus SFU. Los resultados obtenidos con el nuevo lexicón eSOLHotel superan los resultados obtenidos con otro lexicón de propósito general y nos animan a seguir trabajando en esta línea.
Palabras clave: Clasificación de polaridad, corpus de opiniones en español, lexicón dependiente del dominio, turismo.

Abstract: Since Web 2.0 is the largest container for subjective expressions about different topics or issues expressed in all languages, the study of Sentiment Analysis has grown exponentially. In this work, we focus on Spanish polarity classification of hotel reviews and a new domain-dependent lexical resource (eSOLHotel) is presented. This new lexicon has been compiled following a corpus-based approach. We have carried out several experiments using an unsupervised approach for the polarity classification over the category of hotels from corpus SFU. The results obtained with the new lexicon eSOLHotel outperform the results with other general purpose lexicon.
Keywords: Polarity classification, Spanish reviews corpus, dependent-domain lexicon, tourism.

1 Introducción

En los últimos años, el interés por el Análisis de Sentimientos (AS) (conocido en inglés como sentiment analysis u opinion mining) ha crecido significativamente debido a diferentes factores (Pang y Lee, 2008) (Liu, 2012) (Tsytsarau y Palpanas, 2012). Por una parte, el incremento de la creación y compartición de datos por parte de los usuarios de Internet haciendo uso de las nuevas plataformas y servicios que están emergiendo continua y expeditamente. Por otra parte, el consumo de datos online comienza a ser una tarea imprescindible y rutinaria para la toma de decisiones a nivel individual o colectivo.

Muchas son las tareas estudiadas en AS, siendo una de las más consolidadas la clasificación de la polaridad. En esta tarea se han seguido distintas aproximaciones, aunque son dos las líneas principales. Por una parte, la aproximación basada en técnicas de aprendizaje automático (Machine Learning ML), la cual se basa en entrenar unos modelos a partir de una colección de datos etiquetada a priori, con el objetivo de predecir el valor de salida correspondiente a cualquier dato de entrada válido. Los clasificadores pueden estar basados

ISSN 1135-5948

en distintos algoritmos, entre los más utilizados están las máquinas de soporte vectorial (conocido en inglés como Support Vector Machines, SVM) o máxima entropía (ME). Estos clasificadores tienen el inconveniente de necesitar gran cantidad de datos de entrada para un entrenamiento previo y poder obtener buenos resultados. Trabajos como el de Pang, Lee y Vaithyanathan (2002) usan este enfoque supervisado para resolver el problema de la clasificación de polaridad.

La segunda línea, conocida como aproximación basada en Orientación Semántica (OS), obtiene la polaridad de cada documento como la agregación de la inclinación positiva o negativa de sus palabras. La polaridad de las palabras puede ser determinada por diferentes métodos, por ejemplo usando una lista de palabras de opinión (Hu y Liu, 2004), utilizando búsquedas en la web (Hatzivassiloglou y Wiebe, 2000), consultando en una base de datos léxica como WordNet (Kamps et al., 2004) o considerando alguna característica lingüística para determinar el sentimiento a nivel de palabra (Ding y Liu, 2007) (Hatzivassiloglou y Mckeown, 1997) (Turney, 2002). Esta aproximación no necesita de una colección de datos etiquetada a priori para un entrenamiento previo, aunque sí de recursos léxicos normalmente dependientes del idioma para determinar la polaridad de las palabras. Aunque ambas aproximaciones tienen ventajas e inconvenientes, nuestro trabajo se engloba en la aproximación basada en OS. Muchos investigadores han guiado sus pasos intentando resolver estos problemas pero aún quedan otros retos que afrontar y abordar, como es la adaptación de la clasificación de opiniones al dominio tratado (Aue y Gamon, 2005). Es en este reto donde centraremos el esfuerzo de este artículo.

Por otra parte, la mayoría de los trabajos en AS tratan con documentos escritos en inglés a pesar de que cada vez es mayor la cantidad de información subjetiva que publican los usuarios de Internet en su propio idioma. Es por esta razón, que la generación y uso de recursos propios en el idioma de los documentos a tratar se esté convirtiendo en un tema crucial para realizar la clasificación de opiniones mediante orientación semántica. Así pues, nuestro artículo está enfocado al AS en español, de manera que los recursos que utilizaremos estarán en este idioma, tanto corpora como lexicones.

Resumiendo, el desarrollo de recursos lingüísticos nuevos es muy importante para seguir progresando en AS. Además, se hace necesario que esos nuevos recursos se implementen en otros idiomas distintos al inglés, como el español por ejemplo. Así, la descripción de un corpus nuevo de opiniones en el dominio turístico, la descripción de un lexicón de palabras con sentimientos dependiente del dominio y unos experimentos que certifiquen la validez de dichos recursos son la principal contribución de este artículo.

El presente artículo se estructura de la siguiente manera: en la sección 2 se describen brevemente otros trabajos relacionados con la clasificación de polaridad en opiniones escritas en español, trabajos que generan nuevos recursos léxicos y algunos trabajos relacionados con la adaptación al dominio en AS. En la sección 3 se explican los diferentes recursos utilizados, así como la metodología utilizada para la generación del nuevo lexicón adaptado al dominio. En la sección 4 se muestran los experimentos realizados y se discuten los resultados obtenidos. Por último, se exponen las conclusiones y el trabajo futuro.

2 Trabajos relacionados

Centrándonos en los trabajos realizados sobre AS, a continuación se presentan los más relevantes en un idioma distinto del inglés. Como primer trabajo se tiene el de Banea et al. (2008), el cual propone varios enfoques para el análisis de la subjetividad en varios idiomas mediante la aplicación directa de las traducciones de un corpus de opiniones etiquetadas en inglés para el entrenamiento de un clasificador de opiniones en rumano y español. Este trabajo muestra que la traducción automática es una alternativa viable para la construcción de recursos y herramientas para el análisis de la subjetividad en un idioma distinto al inglés. Brooke et al. (2009) presentan varios experimentos relacionados con recursos en español e inglés. Llegan a la conclusión de que, aunque las técnicas de aprendizaje automático pueden proporcionar un buen rendimiento, es necesario integrar el conocimiento y los recursos específicos del idioma con el fin de lograr una mejora notable. Se proponen tres enfoques: el primero utiliza los recursos de forma manual y automáticamente generados para el español. El segundo aplica aprendizaje automático sobre un corpus español y el último

traduce los corpus del español al inglés y luego aplica SO-CAL, (Semantic Orientation Calculator), una herramienta desarrollada por ellos mismos (Taboada et al., 2011). Martínez-Cámara et al. (2011) emplean un corpus de críticas de cine llamado MuchoCine (Cruz et al., 2008) para clasificar opiniones escritas en español usando un enfoque supervisado, y Martín-Valdivia et al. (2013) empleando el mismo corpus de cine en español y generando el corpus paralelo en inglés MCE realiza una combinación de la clasificación supervisada sobre ambos corpus y una clasificación no supervisada integrando SentiWordNet (Esuli and Sebastiani, 2006) sobre el corpus en inglés.

Para realizar la clasificación de la polaridad siguiendo un enfoque basado en orientación semántica, muchos autores usan o generan recursos léxicos en el idioma en el que están escritas las opiniones. Así, Taboada y Grieve (2004) ponen a disposición de los investigadores el corpus SFU en inglés, con 400 opiniones distribuidas en 8 categorías, con 25 opiniones positivas y otras 25 negativas cada categoría. Al poco tiempo, generan otro corpus en español siguiendo la misma filosofía, con 8 categorías similares, el corpus SFU en español. En Cruz et al. (2008) se describe la generación de un corpus MC de críticas de cine escritas en español a partir de la página web MuchoCine.com[1]. El corpus cuenta con 1.274 opiniones clasificadas como negativas y 1.351 opiniones clasificadas como positivas. Boldrini et al. (2009) presentan el corpus EmotiBlog que incluye comentarios sobre varios temas en tres idiomas: español, inglés e italiano. En Molina-González et al. (2013) se presenta un nuevo recurso para la comunidad investigadora en AS en español. El recurso llamado iSOL, el cual será utilizado en este artículo, es una lista de palabras de opinión generada a partir del conocido y ampliamente usado lexicón existente en inglés de Bing Liu (Hu and Liu, 2004). En Díaz-Rangel et al. (2014) se proporciona un lexicón de emociones en español compuesto de 2.036 palabras que llevan asociado un factor de probabilidad de uso afectivo (PFA) con respecto al menos una de las emociones básicas: alegría, enfado, tristeza, sorpresa y disgusto.

Por otra parte, como es bien sabido, la orientación semántica de muchas palabras es dependiente del dominio que se trate, existiendo diversos documentos que corroboran este hecho como son Engström (2004), Owsley, Sood y Hammond (2006) y Blitzer, Dredze y Pereira (2007). Existen trabajos más actuales como Dehkharghani et al. (2012), en el que se propone un método para construir un sistema de clasificación de la polaridad dependiente del dominio. El dominio seleccionado por los autores es sobre comentarios de los huéspedes de hoteles. Cada opinión se representa por un conjunto de características independientes del dominio y otro conjunto dependiente del dominio. En Demiroz et al. (2012) se propone un método para adaptar un recurso lingüístico de sentimientos independiente del dominio, como SentiWordNet, a un dominio específico. En Molina-González et al. (2013) se detalla la generación de un recurso léxico basado en listas de palabras de opinión adaptado al dominio de cine. Nuestra propuesta sigue un enfoque basado en corpus, pero en este caso el dominio utilizado es el turístico y concretamente, usaremos un corpus con opiniones extraídas de TripAdvisor para diferentes hoteles de Andalucía. Los buenos resultados obtenidos en los experimentos demuestran que nuestra propuesta es válida independientemente del dominio elegido.

3 Recursos: corpora y lexicones

En esta sección se describe, en primer lugar, el corpus de opiniones sobre hoteles. Este corpus se llama COAH (Corpus of Opinions about Andalusian Hotels) y está disponible libremente[2]. Los lexicones usados para la experimentación son el lexicón iSOL, independiente del dominio, usado en varios trabajos como Molina-González et al. (2013) y el nuevo lexicón eSOLHotel (iSOL enriquecido para el dominio de hoteles) generado a partir del corpus COAH. El corpus usado para probar la bondad del lexicón generado eSOLHotel es el corpus SFU en español[3], en particular, las opiniones pertenecientes a la categoría de hoteles.

3.1 Corpus COAH

Para compilar un corpus de opiniones es muy importante saber elegir la fuente de dichos

[1] http://www.muchocine.net/

[2] http://sinai.ujaen.es/coah

[3] https://www.sfu.ca/~mtaboada/download/downloadCorpusSpa.html

datos. En nuestro caso, hemos intentado satisfacer los siguientes requisitos:

- Debe haber gran cantidad de opiniones y éstas deben ser escritas por usuarios de los hoteles.
- Cada opinión debe estar valorada por el propietario de dicha opinión.
- El portal web debe ser un portal confiable en el dominio de hoteles.
- Debe ser un portal prestigioso internacionalmente en la búsqueda de información sobre hoteles.

Después de estudiar varios portales web, nuestra elección final fue TripAdvisor[4]. El corpus generado consiste en una colección de opiniones escritas por usuarios no necesariamente profesionales. Este hecho incrementa la dificultad de la tarea, porque los textos pueden no ser gramaticalmente correctos, incluso contener palabras mal escritas o expresiones informales. Se han seleccionado solo hoteles andaluces. Por cada provincia de Andalucía (Almería, Cádiz, Córdoba, Granada, Jaén, Huelva, Málaga and Sevilla), se han elegido 10 hoteles, siendo 5 de ellos de valoración muy alta y los otros 5 con las peores valoraciones, para obtener las mínimas opiniones neutras en el corpus. Todos los hoteles seleccionados deben tener al menos 20 opiniones escritas en español en los últimos años. Finalmente, se han obtenido 1.816 opiniones.

Las opiniones están valoradas en una escala de 1 a 5. El valor 1 significa que el autor manifiesta una opinión muy negativa sobre el hotel, mientras que una puntuación de 5 quiere decir que el autor tiene muy buena opinión sobre el hotel. Los hoteles con valor 3 se pueden catalogar como hoteles neutros, ni buenos ni malos, y por tanto, difíciles de clasificar. En la Tabla 1 se muestra el número de opiniones por valoración.

Valoración	Número de opiniones
1	312
2	199
3	285
4	489
5	531
Total	1.816

Tabla 1: Distribución por valoración

#Opiniones	1.816
#Hoteles	80
Media de opiniones por hotel	22,7
#Palabras	264.303
#Frases	9.952
#Adjetivos	17.800
#Adverbios	15.219
#Verbos	38.590
#Sustantivos	53.640
Media de palabras por frase	26,55
Media de palabras por opinión	145,54
Media de adjetivos por opinión	9,80
Media de adverbios por opinión	8,38
Media de verbos por opinión	21,25
Media de sustantivos por opinión	29,54

Tabla 2: Estadísticas de COAH

En la Tabla 2 se muestran algunas características del corpus. De los metadatos mostrados en la Tabla 2, se puede resaltar que las opiniones tienen una media de 145 palabras suficientes para dar la opinión subjetiva sin implicarse en descripciones objetivas fuera de nuestro estudio. Las páginas web extraídas fueron transformadas en ficheros xml (uno por hotel). Cada fichero xml tiene 20 opiniones. Cada opinión tiene dos tipos de información, una sobre el hotel y otra sobre la opinión del huésped del hotel.

A partir de los ficheros xml se genera un documento que solo alberga la valoración de un hotel específico, el título y la opinión. Para los experimentos se descartan aquellas opiniones neutras, es decir, con valoración 3. El resto de opiniones son catalogadas como positivas si su valoración es 4 ó 5, y negativas si su valoración es 1 ó 2. Por tanto, la clasificación binaria de las opiniones sobre hoteles del corpus COAH es la que se muestra en la Tabla 3.

Clases	Número de opiniones
Positiva	1.020
Negativa	511
Total	1.531

Tabla 3: Clasificación binaria del corpus COAH

En las Figuras 1 y 2 se muestra un ejemplo de un hotel, en XML y en formato texto.

[4] http://www.tripadvisor.es

<ID>1</ID>
<Nombre>Alcazaba Mar Hotel</Nombre>
<Categoria>4</Categoria>
<Dirección>Juegos del Argel, Urbanizacion E1 Toyo |Cabo de Gata </Dirección>
<CódigoPostal>04131</CódigoPostal>
<localidad>Retamar</localidad>
<Provincia>Almería</Provincia>
<País>España</País>
<Viajero>----------</Viajero>
<Localidad_Viajero>-------
</Localidad_Viajero>
<Valoración>3</Valoración>
<Título>"Adecuada la calidad al precio del hotel"</Título>
<Opinión>Acabamos de llegar del hotel. La verdad es que nos fuimos con mucho miedo por los comentarios escritos aquí. Nuestra opinion es que es un hotel comodo, tiene piscina buena, animacion excelente, y un personal muy amable. Quizas lo mas tenido en cuenta es el buffet..... </Opinión>
<Fecha_TipoViajero>Se alojó el Agosto de 2012, viajó con la familia</Fecha_TipoViajero>
<Relación_calidad-precio>3</Relación_calidad-precio>
<Ubicación>2</Ubicación>
<Calidad_del_sueño>3</Calidad_del_sueño>
<Habitaciones>3</Habitaciones>
<Limpieza>3</Limpieza>
<Servicio>4</Servicio>

Figura 1: Ejemplo de un hotel en el corpus COAH

Valoración	Título	Opinión
1	"Un hotel digno de mención!"	Como bien les comenté a los propietarios a la hora de abandonar el hotel, no dudaré un momento en recomendar una y otra vez el Hotel Albero de Granada. Su situación respecto del centro de Granada no es la mejor, pero para nuestros propósitos era perfecto (escapada de fin de semana con visita a la Alhambra). Se encuentra en la carretera de..... ..... Si vuelvo a Granada no dudaré en hospedarme en el mismo hotel. Muchas gracias por todo!!

Figura 2: Fragmento de una opinión del corpus COAH

3.2 Corpus SFU

Para realizar los experimentos, se elige parte del corpus SFU Corpus. El Corpus SFU se compone de opiniones de productos en inglés y español. La versión en inglés (Taboada y Grieve, 2004) tiene 400 opiniones (200 positivas y 200 negativas) de productos comerciales descargados de la web Epinions[5] en el año 2004. Se divide en ocho categorías: libros, coches, ordenadores, utensilios de cocina, hoteles, películas, música y teléfono. Cada categoría incluye 25 opiniones positivas y 25 opiniones negativas. Posteriormente, los autores de SFU Corpus hacen disponible la versión española del corpus[6], con el objetivo de ofrecer un corpus comparable para las siguientes investigaciones. Las opiniones en español se dividen en ocho categorías similares, y también cada categoría tiene 25 opiniones positivas y 25 opiniones negativas. En este caso, las opiniones se descargan desde la web Ciao.es[7]. Para realizar nuestros experimentos se eligen las opiniones de la categoría hoteles.

3.3 Lexicón iSOL

Este recurso fue generado a partir del lexicón en inglés de Bing Liu (Hu y Liu, 2004) traduciéndolo automáticamente al español, obteniendo el recurso SOL (Spanish Opinion Lexicon). Posteriormente, la lista fue revisada manualmente. La lista final de palabras de opinión se llama iSOL (improved SOL). El lexicón iSOL se compone de 2.509 palabras positivas y 5.626 palabras negativas, en total, el lexicón español tiene 8.135 palabras polarizadas. Este recurso fue evaluado satisfactoriamente en Molina-González et al. (2013) usando el corpus MuchoCine (Cruz et al., 2008). Los resultados mostraron que el uso de la lista mejorada de palabras polarizadas puede ser una buena estrategia para la clasificación de polaridad no supervisada.

3.4 Lexicón eSOLHotel

El lexicón iSOL es de propósito general, sin embargo, el AS es una tarea con un cierto grado de interrelación con el dominio tratado. Dentro de los enfoques seguidos para la compilación de un conjunto de palabras de opinión, el más adecuado para obtener términos con carga semántica dependientes del dominio es el que se conoce como el enfoque basado en corpus (Kanayama y Nasukawa, 2006).

Tomando como referencia el lexicón iSOL, se ha generado una lista de palabras de opinión para el dominio de hoteles. Para la generación

[5] http://www.epinions.com/
[6] https://www.sfu.ca/~mtaboada/download/downloadCorpusSpa.html
[7] http://www.ciao.es/

de la lista de palabras de opinión se ha seguido el enfoque basado en corpus. El elemento clave del enfoque basado en corpus es el uso de una colección de documentos etiquetados según su polaridad. El corpus español seleccionado para el proceso es COAH. Hemos seguido el mismo supuesto que Du et al. (2010), es decir, una palabra debe ser positiva (o negativa) si aparece en muchos documentos positivos (o negativos). Por lo tanto, hemos calculado la frecuencia de la palabra en cada clase de documentos (positivos y negativos). De una manera manual y subjetiva, se han seleccionado 166 palabras positivas y 131 palabras negativas, que cumplen los requisitos de aparecer en una clase más veces que en la otra y tener una orientación positiva o negativa. Por lo tanto, se añadieron las 297 palabras más frecuentes que aún no figuraban en la lista iSOL a la lista final obteniendo un total de 8.432 palabras indicadoras de opinión (2.675 positivas y 5.757 negativas). Esta nueva lista de integración de la información del corpus ha sido llamada eSOLHotel (SOL enriquecido y adaptado al dominio Hotel). En la siguiente Tabla 4 se muestran algunas de las palabras que han sido añadidas.

Palabras positivas	Palabras negativas
ensueño	asqueroso
luminoso	cucaracha
coqueto	desconchones
comodisima	humedades
intachable	mejorable
remodelado	reclamaciones
pasada	tugurio
supercentrico	zulo

Tabla 4: Palabras positivas y negativas añadidas al lexicón eSOLHotel

4 Experimentos y resultados

Antes de llevar a cabo los experimentos, a las opiniones de hoteles del corpus SFU se les ha realizado un *preprocesamiento* con el fin de tener en cuenta los mismos criterios que se han utilizado en la generación de los lexicones iSOL y eSOLHotel. Por ejemplo, las letras mayúsculas se han cambiado a minúsculas, a las vocales acentuadas se les ha quitado el acento y los caracteres especiales han sido separados de las palabras, para aislar dichas palabras.

Para decidir si una opinión se considera positiva o negativa, seguimos un simple método basado en la cuenta del número de palabras incluidas en las listas iSOL y eSOLHotel encontradas en las opiniones de hoteles del corpus SFU etiquetado en español. Así, nuestro método clasifica la opinión como positiva si el número de palabras positivas encontradas es igual o mayor que el número de palabras negativas encontradas, o como negativa en el resto de casos.

En la Tabla 5 se muestran los resultados obtenidos en la categoría de hoteles del corpus SFU en español usando los lexicones iSOL (independiente del dominio) y eSOLHotel (adaptado al dominio de hoteles).

Lexicón	Precisión	Valor F1	Exactitud
iSOL	77,41%	73,52%	70,0%
eSOLHotel	84,72%	81,22%	78,0%

Tabla 5: Resultados obtenidos en la clasificación binaria de corpus SFU usando iSOL y eSOLHotel

Los resultados que se muestran en la Tabla 5 confirman nuestra hipótesis de partida, es decir, que la inclusión de información del dominio en una lista de palabras de opinión genérica mejora los resultados de la clasificación de la polaridad. El porcentaje de mejora en la exactitud que se ha obtenido con la inclusión de información del dominio ha sido de un 11,43%. Siguiendo una metodología muy simple, como la que se ha descrito, se ha obtenido una mejora muy importante.

Con el fin de profundizar en el estudio de la bondad de la metodología seguida para la inclusión de información del dominio, se ha construido un clasificador supervisado. Para ello, se ha aplicado a los documentos un algoritmo de normalización morfológica basado en la eliminación de prefijos y sufijos, lo que en el ámbito del Procesamiento del Lenguaje Natural se conoce como *stemmer*. El algoritmo de *stemming* empleado ha sido el de Porter para español. Tras este proceso, los documentos se han representado como vectores de *unigramas* ponderados por el índice de relevancia TF-IDF. Por tanto, las características que recibirá como entrada el algoritmo de aprendizaje automático serán únicamente el valor TF-IDF de los *unigramas* de los documentos. Por último, se ha realizado una validación cruzada con el

algoritmo SVM. Los resultados que se han obtenido son un 82% y un 82,71% de exactitud y valor F1 respectivamente. De nuevo, los resultados de la Tabla 5 indican la bondad de la metodología presentada en el artículo, dado que la diferencia de valor F1 entre SVM y eSOLHotel es solo de un 1,83%. Por lo tanto, la pérdida de exactitud es tan mínima que puede considerarse aconsejable el uso de la lista en lugar del método supervisado, ya que en este caso no se necesitaría de un modelo de aprendizaje automático previamente entrenado.

5 Conclusiones y trabajos futuros

En este artículo se ha presentado una metodología de adaptación de un lexicón de palabras de opinión a un dominio concreto. Para ello se ha tomado un corpus de opiniones de hoteles como referencia (COAH), se han calculado la frecuencia de los términos que componen el corpus y se han seleccionado las palabras de opinión más representativas del corpus. La metodología se ha evaluado con las opiniones de hoteles del corpus SFU en español. Los resultados que se han obtenido (Tabla 5) ponen de manifiesto la bondad de la metodología y nos animan a seguir perfeccionando la metodología de adaptación al dominio.

El sistema de clasificación se puede todavía mejorar aún más. Como trabajo futuro se va a incluir un tratamiento de la negación basado en reglas lingüísticas específico para español. Este nuevo elemento del sistema nos va a permitir clasificar correctamente las opiniones negativas expresadas con términos positivos negados.

Agradecimientos

Esta investigación ha sido parcialmente financiada por el Fondo Europeo de Desarrollo Regional (FEDER), el proyecto ATTOS (TIN2012-38536-C03-0) del Gobierno de España y el proyecto AORESCU (P11-TIC-7684 MO) del gobierno autonómico de la Junta de Andalucía. Por último, el proyecto CEATIC (CEATIC-2013-01) de la Universidad de Jaén también ha financiado parcialmente este artículo.

Bibliografía

Aue, A. y M. Gamon. 2005. Customizing sentiment classifiers to new domains: A case study. En *Proceedings of Recent Advances in Natural Language Processing (RANLP)*.

Banea, C., R. Mihalcea, J. Wiebe, y S. Hassan, S. 2008. Multilingual subjectivity analysis using machine translation. En *Proc. of the conference on empirical methods in natural language processing*, páginas 127–135. ACL.

Blitzer, J., M. Dredze, y F. Pereira. 2007. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. En *Proceedings of the Association for Computational Linguistics (ACL)*.

Boldrini, E., A. Balahur, P. Martínez-Barco, y A. Montoyo. 2009. Emotiblog: an annotation scheme for emotion detection and analysis in non-traditional textual genres. En *DMIN*, páginas 491–497. CSREA Press.

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. ACL.

Cruz, F.L., J.A. Troyano, F. Enriquez, y J. Ortega. 2008. Clasificación de documentos basada en la opinión: experimentos con un corpus de críticas de cine en español. *Procesamiento de Lenguaje Natural*, Volumen 41, páginas 73-80.

Dehkharghani, R., B. Yanikoglu, D. Tapucu, y Y. Saygin. 2012. Adaptation and use of subjectivity lexicons for domain dependent sentiment classification. En *Data Mining Workshops, 2012 IEEE 12th International Conference on*, páginas 669-673.

Demiroz, G., B. Yanikoglu, D. Tapucu, y Y. Saygin. 2012. Learning domain-specific polarity lexicons. En *Data Mining Workshops, 2012 IEEE 12th International Conference on*, páginas 674-679.

Díaz Rangel, I., G. Sidorov y S. Suárez-Guerra. 2014. Creación y evaluación de un diccionario marcado con emociones y ponderado para el español. *Onomazein, 29, 23 p*

Ding, X. y B. Liu. 2007. The utility of linguistic rules in opinion mining. En *Proc. of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, páginas 811–812.

Du, W.T., S. Cheng, y X. Yun. 2010. Adapting information bottleneck method for automatic construction of domain-oriented sentiment lexicon. En *Proc. of ACM International Conference on Web search and data mining.*

Engström, C. 2004. Topic dependence in sentiment classification. *Master's thesis*, University of Cambridge.

Esuli, A. y F. Sebastiani. 2006. SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. En *Proceedings of Language Resources and Evaluation (LREC).*

Hatzivassiloglou, V. y K. McKeown. 1997. Predicting the semantic orientation of adjectives. En *Proceedings of the eighth conference on European chapter of the association for computational linguistics*, páginas 174–181.

Hatzivassiloglou, V. y J. Wiebe, J. 2000. Effects of adjective orientation and gradability on sentence subjectivity. En *Proceedings of the international conference on computational linguistics (COLING)*, páginas 299–305.

Hu, M. y B. Liu. 2004. Mining and summarizing customer reviews. En *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Seattle, Washington, USA, páginas 168-177.

Kamps, J., M. Marx, R.J. Mokken, y M. de Rijke. 2004. Using WordNet to measure semantic orientations of adjectives. En *LREC, European Language Resources Association.*

Kanayama, H. y T. Nasukawa. 2006. Fully automatic lexicon expansion for domain-oriented sentiment analysis. En *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, páginas 355–363. ACL.

Liu, B. 2012. Sentiment analysis and opinion mining. synthesis lectures on human language technologies. *Morgan and Claypool Publishers.*

Martín-Valdivia, M.T., E. Martínez-Cámara, J.M. Perea-Ortega, y L.A. Ureña-López. 2013. Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches. *Expert Systems with Applications*, 40 (10), páginas 3934–3942.

Martínez-Cámara, E., M.T. Martín-Valdivia, y L.A. Ureña-López. 2011. Opinion classification techniques applied to a Spanish corpus. *Proceedings of the 16th international conference on Natural language processing and information systems, NLDB'11*, páginas 169–176.

Molina-González, M. D., E. Martínez-Cámara, M.T. Martín-Valdivia, y J.M. Perea-Ortega. 2013. Semantic Orientation for Polarity Classification in Spanish Reviews. *Expert Systems with Applications;* 40(18), páginas 7250-7257.

Owsley, S., S. Sood, y K.J. Hammond. 2006. Domain specific affective classification of documents. En *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, páginas 181–183.

Pang, B. y L. Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*(1–2), 1–135.

Pang, B., L. Lee, y S. Vaithyanathan. 2002. Thums up?: Sentiment Analysis classification using machine learning techniques. En *Proceedings of the ACL-02 Conference on Empirical methods in natural language processing. Stroudsburg, PA, USA: Association for Computational Linguistics;* 10:79-86.

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

Taboada, M. y J. Grieve 2004. Analyzing appraisal automatically. *Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text,* páginas 158 - 161. Stanford University, CA.

Tsytsarau, M. y T. Palpanas. 2012. Survey on mining subjective data on the web. *Data Mining and Knowledge. Discovery*. 24, 3 478-514.

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, Stroudsburg, PA, USA*, páginas 417-424.