Clasificación de documentos basada en la opinión: experimentos con un corpus de críticas de cine en español

Experiments in sentiment classification of movie reviews in Spanish

Fermín L. Cruz, Jose A. Troyano, Fernando Enriquez, Javier Ortega
Universidad de Sevilla
Av.Reina Mercedes s/n Sevilla
{fcruz,troyano,fenros,javierortega}@us.es

Resumen: En los últimos años se ha producido un creciente interés por el procesamiento automático de las opiniones contenidas en documentos de texto, en parte como consecuencia del aumento exponencial de contenidos generados por usuarios en la Web 2.0, y por el interés entre otros de empresas y gobiernos en analizar, filtrar o detectar automáticamente las opiniones vertidas por sus clientes o ciudadanos. Tomando como punto de partida trabajos de otros autores para el inglés, en el presente artículo exponemos los resultados obtenidos en la experimentación con un clasificador no supervisado de documentos basado en la opinión para el español. Proponemos también una versión supervisada del clasificador que obtiene un resultado sensiblemente mejor. Como paso previo a la experimentación, y ante la ausencia de recursos en español para desarrollar nuestro trabajo, presentamos un corpus de críticas de cine en español, que ha sido puesto a disposición de la comunidad científica.
Palabras clave: Clasificación de documentos basada en la opinión, orientación semántica, construcción de corpus

Abstract: In recent years, automatic processing of opinions in text documents has received a growing interest. Some possible causes are the exponential increase of user-generated contents in Web 2.0, and also the interest of companies and governments in automatically analysing, filtering or detecting opinions from their customers or citizens. On the basis of some similar works in English by other authors, in this paper we expose the results obtained in the experimentation with an unsupervised sentiment classifier for Spanish. We also propose a supervised version of the classifier that shows a significatively better performance. Experiments have been carried out using a corpus that we have extracted from a web of movie reviews in Spanish. We have made this corpus available to the research community.
Keywords: Sentiment analysis, sentiment classification, opinion mining, semantic orientation, corpus building

1. Introducción

Considerada inicialmente una subdisciplina de la tarea de clasificación de documentos, en los últimos años la *clasificación de documentos basada en la opinión* (conocida en inglés bajo los nombres de *sentiment classification*, *sentiment analysis* o *opinion mining*) ha sido objeto de un creciente interés por parte de la comunidad de investigadores del procesamiento del lenguaje natural. Si en la tarea de clasificación de documentos clásica el problema consiste en decidir la temática de un documento de entre un conjunto de temáticas posibles (por ejemplo, centrándonos en el ámbito de las noticias periodísticas, distinguir cuando nos encontramos ante un texto de política, sociedad o deportes), en la clasificación basada en la opinión se trata de determinar si en el texto se expresan opiniones negativas o positivas. Es desde este prisma, considerando "opinión negativa" y "opinión positiva" como las dos clases de salida de la tarea, bajo el que se considera que la clasificación basada en la opinión es una subdisciplina de la clasificación de documentos.

Sin embargo, la naturaleza subjetiva de los documentos con los que se trabaja (análisis de productos, críticas de cine o música, intervenciones políticas, contenidos genera-

dos por internautas como blogs o foros,...) añaden dificultad a la tarea y hacen necesario plantear soluciones distintas a las utilizadas en la clasificación de documentos clásica (Pang, Lee, y Vaithyanathan, 2002). En la clasificación basada en la opinión entran en juego fenómenos del lenguaje no sólo léxicos, sintácticos y semánticos, sino pragmáticos y en gran medida de conocimiento del mundo. Por ejemplo, a la hora de determinar la polaridad de la opinión "*En esta película el director nos regala otra de las joyas a las que nos tiene acostumbrados*", hay que considerar cuestiones cómo *¿de qué director estamos hablando?*, *¿qué otras películas ha hecho el director?*, o *¿son buenas esas películas?*; sólo en base a este conocimiento previo (expresado o no en el mismo documento que la expresión anterior) podremos decidir si la opinión es positiva, tal como parece desprenderse de la semántica, o si se trata de una opinión negativa con una carga considerable de ironía.

Como suele ocurrir en los primeros años de investigación en una nueva tarea de procesamiento del lenguaje natural, los trabajos publicados hasta la fecha se centran exclusivamente en el inglés. En el presente artículo describimos los primeros pasos que hemos dado para adentrarnos en el problema de la clasificación basada en la opinión de textos en español. Nuestro interés se centra principalmente en reproducir los experimentos iniciales llevados a cabo por *Peter D. Turney* ((Turney, 2002), (Turney y Littman, 2003)) aplicándolos a un corpus de críticas de cine en español de construcción propia. Propondremos también una versión supervisada del clasificador que mejora significativamente la precisión.

El resto del presente artículo se estructura como sigue: en la siguiente sección se introducen algunos trabajos que relacionados con el concepto de *orientación semántica* que es clave en este trabajo. En la tercera sección describimos el proceso de construcción y las características del corpus de críticas de cine que hemos utilizado en nuestros experimentos. En la sección siguiente describimos la tarea y la arquitectura del clasificador, basándonos en (Turney, 2002) y proponiendo algunas variaciones. En la quinta sección presentamos los resultados obtenidos en los experimentos que hemos realizado. Finalmente en la última sección resumimos las conclusiones principales que alcanzamos de nuestras primeras experiencias en el campo de la clasificación de documentos basada en la opinión para el español, y planteamos algunas líneas de trabajo futuro.

2. Antecedentes

El concepto central que utilizaremos en nuestros experimentos será el de *orientación semántica*. La orientación semántica de una palabra o conjunto de palabras (que a partir de ahora llamaremos *término*) se define como un valor real que siendo positivo indica que el término en cuestión tiene implicaciones subjetivas positivas (opinión favorable), y siendo negativo indica lo contrario. Distintos valores absolutos de la medida informan además sobre distintos grados de intensidad en dichas implicaciones. Los primeros intentos de clasificar adjetivos según su orientación semántica de manera automática fueron llevados a cabo en (Hatzivassiloglou y McKeown, 1997) basándose en las conjunciones entre adjetivos. En (Kamps y Marx, 2002) se utilizan las distancias semánticas en WordNet(Fellbaum, 1998) entre la palabra cuya orientación semántica se desea conocer y las palabras *good* y *bad*.

En (Turney, 2002) se describe un clasificador no supervisado basado en la opinión. Dicho clasificador decide el carácter positivo o negativo de un documento en base a la orientación semántica de los términos que aparecen en el mismo. La orientación semántica se calcula mediante el algoritmo *PMI-IR* que será descrito más adelante. Con un planteamiento relativamente simple, este sistema clasifica correctamente un 84 % de los documentos de un pequeño corpus de análisis de coches utilizado por el autor. Sin embargo, cuando se tratan de clasificar críticas de cine, la precisión obtenida es sólo del **65 %**, de lo que se deduce que la clasificación de críticas de cine es una tarea especialmente difícil.

Tomaremos como punto de partida de nuestros experimentos este artículo, tratando en primer lugar de reproducir el clasificador descrito adaptándolo al español.

3. Un corpus para la clasificación basada en la opinión en español

Para poder experimentar la clasificación basada en la opinión en el español, el primer paso es localizar un recurso adecuado.

No nos consta la existencia de ningún recurso de la naturaleza requerida para el español, de forma que nos planteamos la generación del mismo. Tras decidir que nuestros esfuerzos se centrarían en la clasificación de críticas de cine, buscamos alguna web que se dedicara al tema de la que extraer el corpus de forma automática. Las características que debía cumplir nuestra elección eran:

- Un número alto de críticas disponibles (a partir de 2.000).
- En el caso de ser los contenidos generados por los usuarios, asegurarnos una mínima calidad de los textos.
- Cada crítica debe llevar asociada la puntuación que el autor le da a la película en cuestión, lo que nos permitirá distinguir si una crítica contiene una opinión favorable o desfavorable.
- La licencia de publicación de la web debe permitirnos utilizar los contenidos libremente.

Bajo estas condiciones la web elegida fue *Muchocine*[1].

3.1. Construcción del corpus

El primer paso para la construcción del corpus fue la extracción de las páginas html de cada una de las críticas contenidas en Muchocine[2], con fecha de febrero de 2008. Las críticas de cine contenidas en esta web son introducidas en la misma por usuarios y no por críticos especializados. Esto añade un punto de dificultad a la tarea que nos ocupa, puesto que los textos pueden contener faltas de ortografía, incoherencias entre lo que se relata y la puntuación final asignada, divergencia entre los tamaños de las distintas críticas,....

Las páginas HTML extraídas de dicha web fueron transformadas en ficheros XML (uno por cada crítica), en los que además del texto de la misma constan el nombre del autor de la crítica, el nombre de la película comentada, la puntuación asignada (valores de 1 a 5) y un pequeño resumen de la crítica a modo de titular escrito también por el autor. Cada crítica ha sido procesada con la herramienta FreeLing (Atserias et al., 2006) para *tokenizar* y separar en oraciones el texto además de para obtener ficheros adicionales con información léxica, sintáctica y semántica. Toda esta información adicional ha sido almacenada como parte del corpus: lexemas, etiquetas morfosintácticas, árboles de dependencias y *synsets* de WordNet (Fellbaum, 1998). El corpus obtenido tiene un total de 3.878 críticas y aproximadamente 2 millones de palabras, con una media de 546 palabras por crítica. La distribución según puntuaciones es la que se muestra en el cuadro 1.

El corpus está disponible[3] para su utilización libre por parte de aquellos investigadores que deseen realizar experimentos de clasificación de documentos basada en la opinión en español.

Puntuación	**Nº de críticas**
1	351
2	923
3	1.253
4	890
5	461
Total	3.878

Cuadro 1: Distribución según puntuaciones del corpus

4. Clasificando documentos de opinión en español

En esta sección describimos el proceso de clasificación de documentos basada en la opinión que hemos utilizado en los experimentos con críticas de cine. Un documento de opinión será cualquier unidad de texto en la que se recoja un análisis crítico sobre algún objeto, pudiendo ser ese objeto un producto comercial, una película, una ley o cualquier otra entidad susceptible de ser sometida a crítica.

4.1. Definición de la tarea

Sea $D = \{d_1, d_2, \ldots, d_n\}$ un conjunto de documentos de opinión. Sean $C = \{negativa, positiva\}$ las clases de salida del clasificador. La tarea consiste en asignar a cada uno de los documentos de D una clase de C, según el carácter negativo o positivo de las opiniones vertidas en cada documento.

La tarea exige ciertas simplificaciones sobre la naturaleza de los documentos considerados. Por ejemplo, se supone que todas las

[1] www.muchocine.net

[2] Los contenidos extraídos de MuchoCine han sido utilizados bajo las condiciones de la licencia **Creative Commons** con la que están publicados. (*http://creativecommons.org/licenses/by/2.1/es/).*

[3] http://www.lsi.us.es/~fermin/corpusCine.zip

opiniones contenidas en un documento son inequivocamente negativas o positivas, a lo largo de todo el documento, y que todas las opiniones se refieren a un mismo objeto de análisis. Por supuesto, en la práctica esto no ocurre, lo que dificulta la tarea.

4.2. Arquitectura del clasificador

El proceso de clasificación expuesto en (Turney, 2002) es, según el autor, completamente no supervisado, al no precisar de una etapa de entrenamiento. Nosotros creemos que la utilización que se hace en el sistema de búsquedas en la web (ver sección *Cálculo de la orientación semántica*) es en cierto modo un recurso externo que si bien no es utilizado en un proceso de entrenamiento como tal sino directamente en la clasificación, debería al menos matizar el carácter *no supervisado* del sistema.

El algoritmo propuesto de clasificación es el siguiente:

- Dada una crítica, extraer bigramas utilizando una serie de patrones morfosintácticos simples. Se postula que este conjunto contiene al menos algunos bigramas que expresan opinión (y tambien muchos otros bigramas que no indican opinión).
- Para cada uno de los bigramas extraídos, calcular la *orientación semántica* (valor real, positivo o negativo) mediante el algoritmo *PMI-IR (Pointwise Mutual Information - Information Retrieval).*
- A partir de la suma de las orientaciones semánticas obtenidas, clasificar la crítica como positiva si el valor calculado es mayor o igual que cero, y negativa en caso contrario.

En las siguientes secciones detallamos como se lleva a cabo cada uno de los pasos, discutimos posibles debilidades del sistema y planteamos algunas modificaciones posibles.

4.2.1. Extracción de bigramas

El primero de los pasos consiste en extraer un conjunto de bigramas del texto de la crítica. Este paso es fundamental puesto que será sobre estos bigramas sobre los que se calcularán las orientaciones semánticas que determinarán la clase de salida del clasificador para la crítica en proceso. Para extraer los bigramas, Turney utiliza cinco patrones morfosintácticos que hemos modificado ligeramente para adaptarlos a la sintaxis del español. Estos patrones indican categorías morfosintácticas de los bigramas a extraer, y plantean restricciones a la categoría morfosintáctica de la palabra que sucede a dichos bigramas. Los patrones utilizados pueden verse en el cuadro 2.

	Primera palabra	**Segunda palabra**	**Tercera palabra (no se extrae)**
1.	adjetivo	nombre	cualquiera
2.	nombre	adjetivo	no es nombre
3.	adverbio	adjetivo	no es nombre
4.	adverbio	verbo	cualquiera
5.	verbo	adverbio	cualquiera

Cuadro 2: Patrones morfosintácticos para la extracción de bigramas.

Dos son las deficiencias más importantes de este método de extracción de los bigramas a nuestro modo de ver. La primera es que algunas relaciones sintácticas no serán capturadas por estos patrones, en cuanto que las dos unidades léxicas relacionadas no aparezcan una seguida de la otra en el texto. Esto podría solucionarse mediante el uso de patrones basados en los árboles de dependencia. A pesar de que en un primer momento implementamos en nuestros experimentos esta idea, la hemos tenido que descartar debido a la poca calidad de los árboles de dependencias con los que contamos. Téngase en cuenta que trabajamos sobre texto espontáneo, escrito por usuarios no profesionales, y que contiene en múltiples ocasiones frases con dudosa construcción gramatical, faltas de ortografía y otras peculiaridades que dificultan la tarea de análisis sintáctico.

En segundo lugar, muchos de los bigramas que se extraen no están indicando opinión alguna (se mostrarán algunos ejemplos en la sección *Experimentos*). Una primera etapa de clasificación de oraciones en objetivas/subjetivas podría ayudar a solucionar este problema; pero esto sería en sí mismo materia suficiente para otra línea de investigación que por ahora no abordaremos.

4.2.2. Cálculo de la orientación semántica

Para calcular la orientación semántica se utiliza el algoritmo *PMI-IR*, que consiste en estimar la *Información Mutua Puntual* (*Pointwise Mutual Information*) entre el término en cuestión y un par de palabras se-

milla que sirven de representantes inequívocos de orientación semántica positiva y negativa, haciendo uso de un buscador de páginas web para llevar a cabo dicha estimación. La *Información Mutua Puntual* se define entre dos palabras w_1 y w_2 y mide estadísticamente la información que obtenemos sobre la posible aparición de un término a partir de la aparición de otro término:

$$PMI(w_1, w_2) = log_2 \left(\frac{p(w_1 \& w_2)}{p(w_1)p(w_2)} \right)$$

A partir de esta medida estadística la orientación semántica de un término t se calcula de la siguiente manera:

$$SO(t) = PMI(t, excellent) - PMI(t, poor)$$

Para estimar la medida *PMI* se utilizan búsquedas en la web, de manera que la probabilidad de co-aparición de dos términos que consta en el numerador de la fórmula de *PMI* se aproxima mediante el número de páginas web en las que ambos términos aparecen uno cercano al otro. Tras algunas transformaciones algebráicas, la fórmula final para el cálculo de la orientación semántica de un término ($SO(t)$) propuesta por Turney es la siguiente:

$$log_2 \left(\frac{hits(t\ NEAR\ excellent) hits(poor)}{hits(t\ NEAR\ poor) hits(excellent)} \right)$$

, donde $hits(t)$ indica el número de páginas devueltas por el buscador *AltaVista*[4] al buscar t. El operador *NEAR* de *AltaVista* (operador no disponible en otros buscadores como *Google*, razón por la que nos decantamos por utilizar *AltaVista*) devuelve las páginas en las que ambos términos aparezcan en una misma página y a una distancia máxima de 10 palabras. Al número de páginas obtenido se le suma 0,01 para evitar una posible división por cero.

De manera intuitiva, la idea detrás de este cálculo de la orientación semántica es que expresiones que indiquen una opinión positiva aparecerán con mayor frecuencia cerca de una palabra con claras connotaciones positivas como *excellent* y con mucho menor frecuencia cerca de una palabra con connotaciones negativas como *poor*.

[4] www.altavista.com

La adaptación del algoritmo *PMI-IR* al español se reduce a escoger dos semillas apropiadas para el español. Las semillas escogidas han sido *excelente* y *malo*.

4.2.3. Utilización de multiples semillas para el cálculo de la orientación semántica

En un artículo posterior del mismo autor se plantea la utilización de dos conjuntos de semillas positivas y negativas en lugar de una sola semilla de cada tipo (Turney y Littman, 2003). También hemos realizado experimentos utilizando un conjunto de semillas en lugar de una palabra aislada. El conjunto de semillas positivas y negativas utilizado ha sido el siguiente:

- **Positivas**: excelente, buenísimo, buenísima, superior, extraordinario, extraordinaria, magnífico, magnífica, exquisito, exquisita
- **Negativas**: malo, mala, pésimo, pésima, deplorable, detestable, atroz, fatal

En la sección *Experimentos* se contrastan los valores obtenidos para la orientación semántica de algunos términos de ejemplo ya sea utilizando una sola semilla por categoría o usando los conjuntos recién expuestos.

4.2.4. Algoritmo de clasificación

Una vez se han calculado las orientaciones semánticas de todos los bigramas extraídos, el proceso de clasificación propuesto por *Turney* se basa en sumar todos los valores obtenidos y clasificar como positiva la crítica si el resultado obtenido es mayor o igual a cero, y como negativa en caso contrario.

En nuestros experimentos, además de este sencillo acercamiento, hemos implementado una solución alternativa supervisada, consistente en calcular un valor óptimo a utilizar como umbral positivo a partir de un conjunto de críticas de entrenamiento. La idea es encontrar un valor real u que maximice el número de críticas positivas del corpus de entrenamiento que obtienen un valor total de orientación semántica mayor o igual que u y el número de críticas negativas que obtienen un valor total de orientación semántica menor que u. Una vez calculado este valor, una crítica será clasificada como positiva si el valor total de la orientación semántica iguala o supera dicho valor, y como negativa en caso contrario.

5. *Experimentos*

Para llevar a cabo algunos experimentos de clasificación, seleccionaremos aleatoriamente 400 críticas de nuestro corpus. De éstas, 200 tienen una puntuación de 1 o de 2, y serán consideradas críticas negativas. Las otras 200 tienen una puntuación de 4 o 5, y serán consideradas críticas positivas. No hemos utilizado el total de las críticas incluidas en el corpus por limitaciones de tiempo: el cálculo de la orientación semántica es un proceso lento al depender del buscador *AltaVista* (entre cada solicitud al buscador hemos de dejar cinco segundos de espera para no saturar al servidor de AltaVista, que de otra manera nos denegaría el servicio). Creemos de todas formas que las 400 críticas que nos servirán para obtener resultados son suficientes teniendo en cuenta que en el artículo de Turney (Turney, 2002) se utilizaban tan solo 120 críticas. Actualmente estamos en proceso de generación de las orientaciones semánticas del resto de las críticas del corpus para utilizarlas en futuros experimentos.

5.1. Resultados en el cálculo de la orientación semántica

En el cuadro 3 se muestran algunos de los bigramas extraídos del corpus utilizado en el proceso de clasificación, y la orientación semántica obtenida usando una sola semilla por clase (SO_{v1}) o varias semillas por clase (SO_{v2}). Los ejemplos incluidos buscan ilustrar algunas de las situaciones que se repiten en el resto del corpus. El último bigrama incluido es un ejemplo de bigrama que no aporta ninguna información acerca de las opiniones vertidas en la crítica.

En general, observamos que la orientación semántica calculada a partir de varias semillas parece funcionar mejor. Por ejemplo, para el bigrama *mínima originalidad* la versión 1 del cálculo de la orientación semántica obtiene un valor ligeramente positivo, lo cuál es erróneo. Esto se ve corregido en la versión 2 del cómputo de la orientación semántica. Existen términos con una orientación semántica a priori ambigüa, como es el caso de *efectos especiales decentes*. En estos casos es dudoso cuál de las versiones de la orientación semántica se comporta mejor.

5.2. Resultados del clasificador no supervisado

En el cuadro 4 se encuentran los resultados obtenidos en la clasificación de las críticas utilizando ambas versiones del cómputo de la orientación semántica, mediante el proceso no supervisado de sumar las orientaciones semánticas obtenidas y clasificar la crítica como positiva si el resultado es mayor o igual que 0 (y negativa en caso contrario). Los resultados obtenidos son comparables a los obtenidos por Turney para el inglés, con una mejora significativa en el caso de utilizar varias semillas para calcular la orientación semántica. Téngase en cuenta además que en el corpus utilizado por Turney las 120 críticas de cine utilizadas correspondían únicamente a dos películas, mientras que en nuestro corpus hay críticas de muchas películas.

5.3. Resultados del clasificador supervisado

La clara desproporción obtenida entre los resultados para las críticas negativas y para las positivas (con semilla simple, 35,5 % para las negativas y 91,5 % para las positivas) nos sugiere la idea de buscar un umbral mejor que 0 para decidir la clase de una crítica a partir de la orientación semántica. En el cuadro 5 se recogen los resultados obtenidos a partir de la clasificación supervisada. Para llevar a cabo la optimización del parámetro *umbral* se ha utilizado un 80 % del corpus de 400 críticas anterior, dejando el 20 % restante para evaluación. Los resultados obtenidos en este segundo acercamiento son significativamente superiores a los obtenidos anteriormente, consiguiendo el sistema clasificar correctamente el 77,5 % de las críticas. Resulta llamativo observar que en la versión supervisada la utilización de una única semilla por clase para el cálculo de las orientaciones semánticas conduce a mejores resultados que el cálculo utilizando varias semillas por clase. Pensamos que en el caso en que usamos una sola semilla, la optimización del parámetro *umbral* compensa una aparente asimetría entre la intensidad de las orientaciones semánticas expresadas por las semillas *malo* y *excelente* (estas semillas son traducciones directas de las utilizadas en inglés por *Turney*). Esta asimetría parece verse mitigada en el caso del uso de múltiples semillas, y es por eso que usando esta última versión del cálculo de la orientación semántica la optimización del

parámetro *umbral* no nos hace mejorar tan espectacularmente los resultados del clasificador.

Término	SO_{v1}	SO_{v2}
mínima originalidad	0,23	-6,08
insufrible sucesión	0,23	-0,37
efectos especiales decentes	-5,08	-7,18
película típica	-0,87	-2,32
estupenda dirección	3,59	2,62
gusto exquisito	7,58	3,37
altamente recomendable	5,49	6,99
fantástico currículum	4,24	0,23
banda sonora	1,61	3,63

Cuadro 3: Orientaciones semánticas de algunos de los bigramas extraídos.

	aciertos positivas	aciertos negativas	aciertos total
SO_{v1}	35,5 %	91,5 %	63,5 %
SO_{v2}	70 %	69 %	69,5 %

Cuadro 4: Resultados para la clasificación no supervisada.

	umbral	aciertos positivas	aciertos negativas	aciertos total
SO_{v1}	13,0	72,5 %	82,5 %	77,5 %
SO_{v2}	-2,25	75 %	72,5 %	73,75 %

Cuadro 5: Resultados para la clasificación supervisada.

6. Conclusiones

En el presente trabajo hemos descrito nuestras primeras experiencias en la clasificación de documentos basados en la opinión para el español. Creemos que dicha tarea y otras relacionadas con el procesamiento automático de las opiniones ofrecen grandes oportunidades de investigación, especialmente aplicadas al español. Los resultados que hemos obtenido, si bien cubren nuestras expectativas como primer acercamiento que hacemos al problema de la clasificación de documentos basada en la opinión, distan aún mucho de lo que cabe esperar de un sistema de clasificación confiable. A la vista del incremento en precisión obtenido en la versión supervisada de nuestro clasificador, creemos que la aplicación de acercamientos supervisados más sofisticados que el aquí propuesto, basados en las soluciones clásicas a la clasificación de documentos pero enriquecidos con la información proporcionada por la orientación semántica, pueden suponer mejoras considerables. Además de esta vía de continuación del presente trabajo, nos planteamos experimentar con otros algoritmos para el cálculo de la orientación semántica (basados en la similitud de palabras y WordNet,como en (Kamps et al., 2004) o (Hu y Liu, 2004); o basados en las frecuencias relativas de aparición en las distintas clases de un corpus , como en (Cane Wing-ki Leung y lai Chung, 2006)), que no dependan de un servicio externo como es *AltaVista*.

Para poder llevar a cabo nuestros experimentos hemos presentado un corpus de críticas de cine en español, que ha sido creado a partir de las críticas introducidas por usuarios en la web *Muchocine*. El corpus de críticas de cine en español está disponible[5] para su utilización libre por parte de aquellos investigadores que deseen realizar experimentos de clasificación de documentos basada en la opinión en español.

Existen multitud de contenidos generados por usuarios en la web que son idóneos para la creación de recursos en los que entrenar o evaluar sistemas relacionados con el procesamiento automático de opiniones. Creemos que la utilización de estos contenidos para la creación de recursos es un punto fundamental para el avance de la investigación en procesamiento automático de opiniones, por lo que será también una de nuestras líneas preferentes de trabajo futuro.

Bibliografía

Atserias, J., B. Casas, E. Comelles, M. González, L. Padró, y M. Padró. 2006. Freeling 1.3: Syntactic and semantic services in an open-source nlp library. En *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, páginas 48–55.

Cane Wing-ki Leung, Stephen Chi-fai Chan y Fu lai Chung. 2006. Integrating collaborative filtering and sentiment analysis: A rating inference approach. En *Proceeding of the ECAI 2006 Workshop on Recommender Systems, in conjunction with*

[5] http://www.lsi.us.es/~fermin/corpusCine.zip

the 17th European Conference on Artificial Intelligence, páginas 62–66.

Fellbaum, Christiane, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, May.

Hatzivassiloglou, Vasileios y Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives. En Philip R. Cohen y Wolfgang Wahlster, editores, *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, páginas 174–181, Somerset, New Jersey. Association for Computational Linguistics.

Hu, Minqing y Bing Liu. 2004. Mining and summarizing customer reviews. En *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, páginas 168–177, New York, NY, USA. ACM.

Kamps, J., M. Marx, R. Mokken, y M. de Rijke. 2004. Using wordnet to measure semantic orientation of adjectives.

Kamps, Jaap y Maarten Marx. 2002. Words with attitude. En *1st International WordNet Conference*, páginas 332–341.

Pang, Bo, Lillian Lee, y Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. En *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Turney, Peter D. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. En *ACL*, páginas 417–424.

Turney, Peter D. y Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Trans. Inf. Syst.*, 21(4):315–346.