Procesamiento del Lenguaje Natural
ISSN: 1135-5948
secretaria.sepln@ujaen.es
Sociedad Española para el Procesamiento del Lenguaje Natural
España

Fernández, Javi; Gómez, José Manuel; Boldrini, Ester; Martínez-Barco, Patricio
Análisis de Sentimientos y Minería de Opiniones: el corpus EmotiBlog
Procesamiento del Lenguaje Natural, núm. 47, septiembre, 2011, pp. 179-187
Sociedad Española para el Procesamiento del Lenguaje Natural
Jaén, España

Disponible en: http://www.redalyc.org/articulo.oa?id=515751747019

Procesamiento del Lenguaje Natural, Revista nº 47 septiembre de 2011, pp 179-187 recibido 02-05-2011 aceptado 23-05-2011

Análisis de Sentimientos y Minería de Opiniones: el corpus EmotiBlog
Sentiment Analysis and Opinion Mining: The EmotiBlog Corpus

Javi Fernández
University of Alicante, Department of Language and Computing Systems
javifm@ua.es

Ester Boldrini
University of Alicante, Department of Language and Computing Systems
eboldrini@dlsi.ua.es

José Manuel Gómez
University of Alicante, Department of Language and Computing Systems
jmgomez@ua.es

Patricio Martínez-Barco
University of Alicante, Department of Language and Computing Systems
patricio@dlsi.ua.es

Resumen*:* EmotiBlog es una colección de entradas de blogs creado y anotado para detectar expresiones subjetivas en los nuevos géneros textuales nacidos con la Web 2.0. Investigaciones previas han demostrado la relevancia de los sistemas de aprendizaje automático como recurso para la detección de información de opinión. En este artículo exploramos características adicionales para un análisis profundo de estas técnicas. Además, comparamos EmotiBlog con la colección JRC. Los resultados obtenidos demuestran la validez de EmotiBlog y nos animan a seguir en esta línea de investigación.
Palabras clave: Análisis de Sentimientos, EmotiBlog, Aprendizaje Automático.

Abstract: EmotiBlog is a collection of blog posts created and annotated for detecting subjective expressions in the new textual genres born with the Web 2.0. Previous work has demonstrated the relevance of the Machine learning systems as tool for detecting opinionated information. In this paper we explore additional features for a deep analysis of these techniques. Moreover, we compare EmotiBlog with the JRC collection. The obtained results demonstrate the usefulness of EmotiBlog and support us to continue in this research path.
Keywords: Sentiment Analysis, EmotiBlog Corpus, Machine Learning.

1 *Introducción y motivación*[1]

Gracias a la expansión de la Web 2.0, los últimos años han visto un crecimiento exponencial de la información subjetiva disponible en Internet. Este fenómeno ha originado interés por el *análisis de sentimientos* (AS), una tarea del *procesamiento del lenguaje natural* (PLN) que identifica opiniones relacionadas con un objeto (Liu, 2006). Los datos subjetivos tienen un gran potencial; de hecho, pueden ser explotados por administraciones públicas, empresas y particulares para, por ejemplo, conocer la opinión de los personalidades públicas, elegir la propaganda idónea según las preferencias u opiniones de la gente o encontrar el producto mejor valorado por los usuarios (Liu, 2007). Dado el tipo de registro más informal que los usuarios suelen utilizar (aunque sea no en la totalidad de los casos), el empleo de técnicas de PLN resulta complicado ya que las herramientas que tenemos a disposición no contemplan irregularidades lingüísticas en la mayoría de los casos. Teniendo en cuenta este contexto, nuestra investigación está motivada principalmente por la carencia de recursos, métodos y herramientas para un efectivo procesamiento de la información subjetiva. Por tanto, el principal objetivo de este artículo es demostrar que el corpus y el modelo de anotación de *EmotiBlog* pueden ser un recurso válido y robusto para contribuir a superar los desafíos del AS. En los experimentos observamos que nuestra aproximación contribuye a resolver la carencia de datos anotados con una granularidad gruesa. Hemos entrenado un modelo usando sistemas de *aprendizaje automático* (AA) con los corpus *EmotiBlog Kyoto*[2] y *EmotiBlog*

[1] Este trabajo ha sido parcialmente financiado por los proyectos TEXTMESS 2.0 (TIN2009-13391-C04-01) y Prometeo (PROMETEO/2009/199), y por la acción complentaria de la Generalitat Valenciana (ACOMP/2011/001).

[2] El corpus *EmotiBlog* está compuesto por entradas de blogs sobre el Protocolo de Kyoto y las Elecciones en Zimbabwe y

Phones[3] y además con la colección *JRC*[4]. Estos experimentos han sido posibles ya que los corpus comparten ciertos elementos anotados comunes (mirar sección 3). De esta forma hemos obtenido un conjunto mayor de resultados comparables. Una vez realizada esta comparación, hemos integrado dos recursos léxicos con relaciones semánticas: *SentiWordNet* (Esuli and Sebastiani, 2006) y *WordNet* (Miller, 1995) para aumentar la cobertura de los resultados sin disminuir la precisión. Para mejorar los resultados de los modelos de aprendizaje supervisado, hemos utilizados técnicas de PLN (*stemming*, *lematización*, *bolsa de palabras*, etc.). En trabajos previos se ha demostrado que *EmotiBlog* es un recurso beneficioso para la *búsqueda de respuestas de opinión* en Balahur et al. (2009c y 2010a,b) o el *resumen automático* de textos subjetivos (Balahur et al. 2009a). Por tanto, el primer objetivo de esta investigación es demostrar que *EmotiBlog* es un recurso útil para entrenar sistemas AA enfocados a varias aplicaciones. Como veremos en la sección 2, la mayor parte de los trabajos hechos en *minería de opiniones* (MO) sólo clasifican la polaridad del sentimiento en positivo o negativo. Por lo tanto, nuestro segundo objetivo, es demostrar que el uso de *EmotiBlog* puede ser beneficioso para valorar otras características como la intensidad y la emoción gracias a su etiquetado con granularidad fina. Así, nuestro tercer objetivo es demostrar cómo una clasificación más profunda de la tarea de MO es crucial (mirar sección 2) para avanzar en esta área de investigación. Creemos que hay una necesidad de determinar la intensidad y el tipo de emoción (Boldrini et al, 2009a) como parte de otros elementos presentados en Boldrini et al. (2010).

2 *Trabajos relacionados*

El AS es una disciplina del PLN que recientemente ha originado interés en la comunidad científica generando algunos recursos como *WordNet Affect* (Strapparava and Vilitutti, 2004), *SentiWordNet* (Esuli and Sebastiani, 2006), *Micro-WNOP* (Cerini et. Al, 2007) – estos dos últimos con relaciones semánticas con *WordNet* – o *Emotion triggers* por Balahur and Montoyo (2008). Estos recursos contienen palabras sueltas cuya polaridad y emoción no son necesariamente aquellas anotadas dentro del recurso en un contexto más amplio. En nuestro trabajo, sin embargo, hemos creado un corpus ampliamente anotado tanto a nivel de sentencia como de elementos individuales dentro de la frase para considerar el contexto y su influencia. El punto de partida de la investigación de la emoción se marcó principalmente por el trabajo de Wiebe (1994) quién estableció los puntos de referencia en la configuración del AS para el reconocimiento del lenguaje orientado a opinión y discriminar éste del lenguaje objetivo. Wiebe propuso un método para etiquetar los corpus dependiendo de estos dos aspectos. Nuestro trabajo tiene en consideración esta inicial distinción pero añadimos un nivel más profundo de anotación de la emoción. Ya que expresiones de emoción también están relacionadas con la opción, trabajos previos también incluyen revisiones y comentarios de usuarios para la clasificación a nivel de documento, usando clasificadores de sentimiento, técnicas de AA (Pang and lee, 2003), puntuación de características (Dave, Lawrence and Pennock, 2003), relaciones sintácticas y otros atributos con SVM (Mullen and Collier, 2004), clasificación de sentimientos considerando escalas de valoración (Pang et al, 2002) y métodos supervisados (Chaovalit and Zhow, 2005). Investigación en clasificación a nivel de documentos incluye clasificación de sentimientos de revisiones (Ng et al. 2006), realimentación de los clientes (Gamon et al. 2005) o experimentos comparativos (Cui et al. 2005). Otros analizan sentimientos a nivel de frase usando técnicas de *bootstrapping* como Riloff and Wiebe (2003), o considerando adjetivos (Hatzivassiloglou and Wiebe, 2000) o buscando fuerzas de opinión (Wilson et al., 2004). Otros trabajos incluyen frases comparativas, extracción de relaciones y características basadas en MO y resumen (Turney, 2002) o emplean recursos léxicos con relaciones semánticas como *SentiWordNet* en Ohana y Tierney (2009) o Abulaish et al. (2009). Todos estos están enfocados en encontrar y clasificar la polaridad de las palabras de opinión (mayoritariamente adjetivos) sin tener en cuenta los modificadores o el contexto. Así, nuestro trabajo presenta el primer paso hacia una comprensión contextual de las raíces del lenguaje de las expresiones de opinión y el desarrollo de un sistema de MO para una aplicación concreta. Co-

en Estados Unidos, pero para este trabajo sólo hemos utilizado la parte del Protocolo de Kyoto

[3] Es una extensión de EmotiBlog con opiniones acerca de teléfonos móviles

[4] http://langtech.jrc.ec.europa.eu/JRC_Resources.html

mo veremos en las siguientes secciones, la inclusión de *EmotiBlog* permite un análisis y procesamiento más detallados. Aparte de este recurso, incluimos otros recursos léxicos y técnicas de PLN para alcanzar un entrenamiento automático más efectivo y con mejores resultados.

3 *Corpus*

El corpus que hemos empleado principalmente es *EmotiBlog*[5] *Kyoto* extendido con una colección de páginas Web sobre teléfonos móviles (*EmotiBlog Phones*). La primera parte (*Kyoto*) es una colección de entradas de blogs en inglés acerca del Protocolo de Kyoto, mientras que la segunda (*Phones*) está compuesta por opiniones sobre teléfonos móviles extraídas de Amazon[6]. El modelo de anotación de *EmotiBlog* contempla la anotación a nivel de documento, frase y elemento (Boldrini et al. 2010), distinguiéndolos entre *objetivos* y *subjetivos*. La lista completa de etiquetas así como la explicación de cada una de ellas está disponible en Boldrini et al. (2009a). Para cada elemento se anotan ciertos atributos comunes: *polaridad*, *grado* (o *intensidad*) y *emoción*. Cabe destacar que se ha detectado un alto porcentaje de coincidencias entre los dos experimentados anotadores encargados de etiquetar esta colección en un trabajo previo (Boldrini et al, 2009a), asegurándonos así de la precisión y fiabilidad del etiquetado. La Tabla 1 presenta el tamaño del corpus en número de frases.

	Subjetivas			Objetivas	Total
	Total	POS	NEG		
EmotiBlog Kyoto	210	62	141	347	557
EmotiBlog Phones	246	198	47	172	418
EmotiBlog Full	456	260	188	519	975
JRC	427	193	234	863	1290

Tabla 1: Tamaño de los corpus en frases

Como corpus a comparar con el *EmotiBlog* usamos *JRC*[7], un conjunto de 1590 citas extraídas automáticamente a partir de noticias y, posteriormente, anotadas las frases que expresaban sentimiento (Balahur et al., 2010c). *JRC* tiene una granularidad gruesa en su etiquetado, es decir, sólo contempla objetividad y polaridad para el sentido general de cada frase. Es por eso que, para nuestro propósito de comparar ambos corpus, utilizamos únicamente estos elementos comunes. Cabe destacar que sólo empleamos aquellas frases de JRC en las que todos los anotadores han coincidido. La Tabla 1 también presenta el tamaño de este corpus en frases.

[5] Disponible mediante petición a los autores
[6] http://www.amazon.co.uk
[7] http://langtech.jrc.ec.europa.eu/JRC_Resources.html

4 *Experimentos del sistema de AA*

Con el fin de demostrar que *EmotiBlog* es un valioso recurso para AA, hemos llevado a cabo una serie de experimentos con diferentes aproximaciones, elementos y recursos. Dichos experimentos han sido evaluados mediante *validación cruzada*.

4.1 EmotiBlog sin información semántica

En este primer paso usamos *EmotiBlog Kyoto* y *EmotiBlog Phones* por separado, y una combinación de ambos (*EmotiBlog Full*). La Tabla 2 presenta las estadísticas de las distintas clasificaciones que evaluamos. Las categorías de los diferentes corpus varían porque no todas ellas han sido encontradas y anotadas en todos los corpus.

	Clasificación	Muestras	Categorías
EmotiBlog Kyoto	Objetividad	556	2
	Polaridad	202	2
	Grado	209	3
	Emoción	209	5
	Obj+Pol	549	3
	Obj+Pol+Grado	549	6
EmotiBlog Phones	Objetividad	416	2
	Polaridad	244	2
	Grado	236	3
	Emoción	243	4
	Obj+Pol	416	3
	Obj+Pol+Grado	408	7
EmotiBlog Full	Objetividad	972	2
	Polaridad	446	2
	Grado	445	3
	Emoción	452	5
	Obj+Pol	965	3
	Obj+Pol+Grado	957	7

Tabla 2: Número de muestras y categorías por clasificación

Podemos observar que clasificar la objetividad o la polaridad es más sencillo que clasificar el grado o la emoción debido al mayor número de categorías de estas dos últimas. Sin embargo, para poder evaluar la polaridad necesitamos evaluar primero la objetividad, con el fin de aplicar la polaridad sólo a las frases subjetivas (en la frases objetivas esto no tiene sentido ya que siempre la polaridad será neutra). La misma situación se aplica al grado: necesitamos determinar primero que una frase es subjetiva y además su polaridad, ya que así podremos discernir si el grado se refiere a una opinión positiva o negativa. Por esta razón hemos decidido combinar clasificaciones, para observar si esta

	Clasificación	word		lemma		stem	
		F-measure	Técnicas	F-measure	Técnicas	F-measure	Técnicas
EmotiBlog Kyoto	**Objetividad**	0.6440	tfidf, chi950	0.6425	tfidfn	**0.6577**	**tfidfn, chi250**
	Polaridad	0.7116	jirsn, ig400	0.6942	tfidf, ig200	**0.7197**	**tfidf, ig500**
	Grado	0.5884	tfidf, ig900	**0.6296**	**tfidf, ig350**	0.6146	tfidfn, ig600
	Emoción	0.4437	tfidfn, ig350	**0.4665**	**jirsn, ig650**	0.4520	jirsn, ig650
	Obj+Pol	0.5914	jirsn, ig600	0.5899	tfidfn, ig750	**0.6064**	**jirsn, ig250**
	Obj+Pol+Grado	0.5612	jirsn	**0.5626**	**jirsn**	0.5433	tfidf, ig700
EmotiBlog Phones	**Objetividad**	0.6200	jirsn, ig900	**0.6405**	**tfidfn, chi500**	0.6368	tfidfn, ig600
	Polaridad	**0.7746**	**tfidf, ig250**	0.7719	tfidfn	0.7516	tfidfn, ig500
	Grado	0.6156	tfidfn	**0.6174**	**jirsn, ig650**	0.6150	tfidf, ig650
	Emoción	0.7555	jirsn, ig450	**0.7828**	**jirsn, ig150**	0.7535	tfidf, ig350
	Obj+Pol	0.5287	tfidf, ig650	**0.5344**	**tfidfn, ig900**	0.5227	tfidf, ig850
	Obj+Pol+Grado	0.4395	tfidf, ig700	0.4424	tfidf	**0.4557**	**tfidf, ig600**
EmotiBlog Full	**Objetividad**	0.5964	jirsn, ig150	0.6080	jirsn, chi100	**0.6229**	**jirsn, ig350**
	Polaridad	0.6109	tfidfn, ig1000	**0.6196**	**tfidf, chi100**	0.6138	tfidf, chi50
	Grado	0.5655	jirsn	0.5526	jirsn	**0.5775**	**jirsn, ig450**
	Emoción	0.5675	jirsn, ig850	**0.5712**	**tfidfn, ig800**	0.5644	jirsn, ig800
	Obj+Pol	0.5332	tfidf	0.5381	tfidf, ig700	**0.5431**	**tfidf**
	Obj+Pol+Grado	0.4794	tfidf, ig700	0.4903	tfidf	**0.4923**	**jirsn**

Tabla 3. Mejores resultados obtenidos y técnicas utilizadas

aproximación mejora los resultados en la evaluación de la polaridad y el grado. Hemos combinado polaridad y objetividad (*Obj+Pol*), con las siguientes categorías resultantes: *objetiva*, *positiva* y *negativa*. También hemos combinado el grado con la objetividad y la polaridad (*Obj+Pol+Grado*), con las siguiente siete categorías resultantes: *objetiva*, *positiva baja*, *positiva media*, *positiva alta*, *negativa baja*, *negativa media* y *negativa alta*. Para este primer paso, empleamos la clásica *bolsa de palabras* (**word**). Para reducir las dimensiones de las muestras también hemos utilizado técnicas de *stemming* (**stem**), *lematización* (**lemma**) y métodos *reducción de dimensionalidad por selección de términos* (RDS). Para RDS hemos comparado dos aproximaciones, *ganancia de información* (**ig**) y *chi square* (**x2**), por reducir la dimensionalidad sustancialmente sin perder efectividad (Yang and Pedersen, 1997). Hemos aplicado estas técnicas con diferentes números de características seleccionadas (**ig50**, **ig100**, …, **ig1000**). Para pesar estas características hemos evaluado las técnicas más comunes: *pesado binario* (**binary**), *tf/idf* (**tfidf**) y *tf/idf normalizado* (**tfidfn**) (Salton and Buckley, 1988). También hemos incluido como técnica de pesado la utilizada por Gómez et al. (2006) en *recuperación de información* (RI) para comprobar su fiabilidad en este nuevo ámbito (**jirs**). En resumen podemos decir que este último pesado es similar a *tf/idf* pero sin tener en cuenta la frecuencia de los términos. También hemos utilizado su versión normalizada (**jirsn**). Como método de aprendizaje supervisado hemos elegido *máquinas de soporte vectorial* (SVM) por su calidad en clasificación de textos (Sebastiani, 2002) y los prometedores resultados obtenidos en estudios previos (Boldrini et al. 2009b). La implementación utilizada ha sido la de Weka [8]con los parámetros por defecto. Debido al gran número de experimentos realizados (aproximadamente 1 millón, debido a todas las combinaciones de técnicas posibles) y parámetros de AA ajustados, en la Tabla 3 presentamos sólo los mejores resultados y la combinación de técnicas para cada uno de ellos.

En general, los mejores resultados han sido obtenidos utilizando *lematización* o *stemming*. El *stemming* funciona mejor cuando el número de características es reducido, mientras que cuando es mayor es más apropiado utilizar *lematización*. Los experimentos que utilizan RDS han obtenido mejores resultados que los que no la utilizan, sin diferencias significativas entre *ig* y *x2*. El mejor número número de características seleccionadas oscila entre 100 y 800, dependiendo del número de clases y muestras de la clasificación: cuantas más haya, más características es necesario seleccionar. Podemos decir también que, en general, cualquier técnica de pesado funciona mejor que el *pesado binario*, aunque los resultados son muy similares independientemente del método utilizado.

También observamos que los resultados obtenidos con los corpus de *Kyoto* y *Phones* separadamente nos dan mejores resultados que la unión de ellos (*Full*). Este resultado era de esperar debido a la especialización de ambos corpus: al modelo de AA es más fácil aprender sobre dominios restringidos.

[8] http://www.cs.waikato.ac.nz/ml/weka/

Como podemos deducir a partir de los experimentos, la evaluación de las clasificaciones de objetividad y polaridad es más fácil debido al bajo número de categorías de cada una de ellas. Los resultados obtenidos son más altos que la clasificación del grado con una media de mejoría de 4.9% y 14.2% respectivamente. Además, una vez se ha detectado la objetividad, determinar la polaridad también es más fácil, aunque el número de muestras para la polaridad sea un 41% más pequeño y ambos tengan el mismo número de categorías. La primera tarea es más compleja que la segunda, porque los vectores de características son más cercanos. Esto significa que hay más ambigüedad en la clasificación de la objetividad que en la polaridad. Términos como "*malo*", "*bueno*", "*excelente*" u "*horrible*" claramente determinan la polaridad de las frases (si no se usa ni negación ni ironía) pero es más difícil encontrar este tipo de palabras para distinguir entre subjetivo u objetivo.

Aunque la combinación de categorías (*Obj+Pol* y *Obj+Pol+Deg*) obtiene peor *f-measure*, esto no significa que esta aproximación no sea adecuada. Para obtener las clasificaciones de polaridad y grado de la Tabla 3 se han preseleccionado previamente sólo las frases subjetivas y, como hemos mencionado antes, esto en una aplicación final no es trivial. En este caso sería necesario detectar previamente las frases subjetivas, posteriormente la polaridad y finalmente el grado, lo que propagaría el error en cada clasificación. Si calculamos la *precisión* (P), en lugar de la *f-measure*, de los mejores experimentos de cada categoría separadamente y obtenemos la precisión final propagando los errores multiplicando sus respectivas precisiones, el resultado no sería tan bueno. Cabe destacar que, para propagar el error sólo tenemos en cuenta la precisión de la clase *subjetiva*, evitando la *objetiva*. Por ejemplo, cuando evaluamos objetividad y polaridad usando el corpus *Full*, obtenemos una precisión de **0,71** y **0,72** respectivamente. Sin embargo, la precisión propagada sería el producto de estos valores (0,51), el cual es un **12%** más bajo que evaluar juntos *Obj+Pol*. Esto se hace más destacado si evaluamos el grado dando una precisión **37%** más baja que de forma separada. En la Tabla 4 se muestran las mejores precisiones con los tres corpus principales. Como podemos observar, la precisión combinando clasificaciones son entre un **8,34%** y un **68,39%** mejores. Cuantas más clasificaciones se combinan, mayor es la mejora. Esto ocurre porque en el caso de clasificaciones separadas, el proceso de AA no tiene información acerca del resto de clasificaciones. Cuando las juntamos, esto ya no ocurre.

	Combinación	Precisión
EB Kyoto	P(Obj) · P(Pol)	0.4352
	P(Obj+Pol)	**0.6113**
	P(Obj) ·P(Pol) · P(Grado)	0.2852
	P(Obj+Pol+Deg)	**0.4571**
EB Phones	P(Obj) · P(Pol)	0.5154
	P(Obj+Pol)	**0.5584**
	P(Obj) ·P(Pol) · P(Grado)	0.3316
	P(Obj+Pol+Deg)	**0.4046**
EB Full	P(Obj) · P(Pol)	0.5090
	P(Obj+Pol)	**0.5771**
	P(Obj) ·P(Pol) · P(Grado)	0.3097
	P(Obj+Pol+Deg)	**0.4912**

Tabla 4: Precisión según combinaciones de categorías

4.2 EmotiBlog con Información Semántica

Para poder comprobar el impacto de incluir relaciones semánticas como características de aprendizaje y reducir la dimensionalidad de éstas, se ha realizado un agrupamiento de características según sus relaciones semánticas. En este punto, el desafío fue la *desambiguación de sentidos de las palabras* (DSP) debido a los pobres resultados que este tipo de sistemas tradicionalmente obtienen en competiciones internacionales (Agirre et al. 2010). Suponemos que seleccionar un sentido erróneo para un término introduciría ruido en la evaluación y reduciría el rendimiento del sistema de AA. Pero si incluimos todos los sentidos de ese término, las técnicas de RDS podrían escoger los sentidos correctos. Para conseguir este propósito, se han utilizado dos recursos léxicos: *WordNet* (WN) y *SentiWordNet* (SWN) mencionados en las secciones 1 y 2. Nuestra decisión de usar estos recursos se debe a que el primero tiene una gran cantidad de relaciones semánticas entre términos en inglés, y el segundo ha demostrado mejorar los resultados de los sistemas de MO (Abulaish et al. 2009). Este último asigna a algunos de los sentidos de WN tres puntuaciones: *positividad*, *negatividad* y *objetividad*. Como SWN sólo contiene los sentidos de WN que tienen información subjetiva, queremos comprobar si expandiendo únicamente mediante esos sentidos podemos mejorar los resultados. Además, queremos añadir las puntuaciones mencionadas en el sistema de AA como nuevos atributos. Por ejemplo, si tenemos un sentido *S* con una positivi-

dad de 0,25 y una negatividad de 0,75, añadiríamos por un lado una característica llamada *S* (con el peso dado por el método de pesado), y por otro dos características más: *S-negativa* y *S-positiva*, con las puntuaciones negativa y positiva respectivamente. Estos experimentos con recursos léxicos se han llevado a cabo en cinco configuraciones diferentes utilizando: sólo sentidos de SWN (**swn**), sólo sentidos de WN (**wn**), sentidos de SWN para términos subjetivos y de WN para el resto (**swn+wn**), sólo sentidos de SWN añadiendo las puntuaciones (**swn+scores**) y sentidos de SWN para términos subjetivos añadiendo las puntuaciones y de WN para el resto (**swn+wn+scores**). Cuando un término no se encuentra en ningún recurso léxico, se utiliza su lema directamente. Además, para resolver la ambigüedad, hemos optado por utilizar dos técnicas: añadir todos los sentidos y dejar que los métodos de RDS se encarguen de desambiguar (los mencionados **swn**, **wn**, **swn+wn**, **swn+scores** y **swn+wn+scores**), pero también añadir únicamente el sentido más frecuente para cada término (**snw1**, **wn1**, **swn1+wn1**, **swn1+scores** y **swn1+wn1+scores**). En la Tabla 5 se presentan los mejores resultados añadiendo la información semántica al proceso de clasificación.

	Clasificación	f-measure	Técnicas
EmotiBlog Kyoto	Objetividad	0.6647	swn+wn+scores, tfidf, chi900
	Polaridad	0.7602	swn1, tfidfn, chi550
	Grado	0.6609	swn1, jirsn, ig550
	Emoción	0.4997	swn, tfidf, chi450
	Obj+Pol	0.5893	swn, tfidfn
	Obj+Pol+Gra	0.5488	swn1+wn1, tfidf
EmotiBlog Phones	Objetividad	0.6405	swn1+wn1+scores, jirsn, ig1000
	Polaridad	0.8093	swn+scores, tfidfn, ig550
	Grado	0.6306	swn1+wn1, tfidfn, ig150
	Emoción	0.8133	swn+wn+scores, jirsn, ig350
	Obj+Pol	0.5447	swn+wn+scores, tfidfn, chi200
	Obj+Pol+Gra	0.4445	swn1, jirsn
EmotiBlog Full	Objetividad	0.6274	swn+wn, jirsn, chi650
	Polaridad	0.6374	swn1+scores, jirsn, chi350
	Grado	0.6101	swn1+wn1+scores, tfidf, ig1000
	Emoción	0.5747	swn+wn+scores, jirsn, ig450
	Obj+Pol	0.5493	swn+wn+scores, tfidf, chi950
	Obj+Pol+Gra	0.4980	swn+wn+scores, jirsn

Tabla 5: Resultados con información semántica

Excepto en unos pocos casos, la información semántica de WN y SWN mejora los resultados finales. Esta mejora puede alcanzar un **7,12%**. A pesar de que sólo hemos mostrado los mejores resultados, esta tendencia se observa en el resto de experimentos: las pruebas realizadas usando relaciones semánticas están en las primeras posiciones. Usar sólo WN no funciona tan bien como utilizar también SWN debido a que éste es un recurso específico para AS que contiene información muy valiosa sobre términos subjetivos. De nuevo, no parece haber diferencias significativas entre las distintas técnicas de pesado (exceptuando el *binario*, que siempre da peores resultados). Es importante mencionar el hecho de que las técnicas de RDS aparecen siempre entre los mejores resultados, con lo cual parece demostrar que estos métodos son apropiados para la desambiguación. Pese a todo, en trabajos futuros queremos realizar más experimentos para intentar afirmarlo con más rotundidad.Podemos observar que los mejores resultados incluyen recursos léxicos. Es más, podemos ver en la Tabla 5 que SWN está presente en todos los mejores resultados, y las puntuaciones de positividad y negatividad aparecen en un 55% de ellos. Además, la utilización de estas puntuaciones está presente en casi todos los mejores resultados para el corpus *Full*. Por lo tanto, esta técnica parece ser mejor para los corpus que no pertenecen a un dominio específico. En nuestros próximos experimentos comprobaremos si esta tendencia continúa utilizando un corpus más grande y que abarque un mayor número de dominios. Los experimentos nos animan a continuar utilizando SWN en este tipo de tareas y encontrar nuevas formas de aprovechar la información subjetiva que proporciona.

4.3 Experimentos con el corpus de JRC

Finalmente, se han aplicado todas las técnicas mencionadas también con el corpus *JRC*. En la Tabla 6 presentamos un resumen con los mejores resultados de estos experimentos.

	Classification	f-measure	Techniques
Word	Objetividad	0.6022	tfidfn, ig950
	Polaridad	0.5163	jirsn
	Obj+Pol	0.5648	tfidfn, ig100
Lemma	Objetividad	0.6049	jirsn
	Polaridad	0.5240	tdidfn, ig800
	Obj+Pol	0.5697	jirs
Stem	Objetividad	0.6066	jirsn
	Polaridad	0.5236	tfidfn, ig450
	Obj+Pol	0.5672	tfidf
WN	Objetividad	**0.6088**	wn1, jirsn, ig650
	Polaridad	**0.5340**	wn1, tfidfn, ig800
	Obj+Pol	**0.5769**	wn1, jirsn, ig700
SWN+WN	Objetividad	0.6054	swn1+wn1, jirsn
	Polaridad	0.5258	swn+wn+scores, jirsn
	Obj+Pol	0.5726	swn1+scores, jirsn

Tabla 6: Experimentos con JRC

Podemos observar en primera instancia que los experimentos añadiendo información semántica, tanto WN como SWN, obtienen mejores valores que los experimentos sin ellos. En particular, usan-

do sólo WN obtenemos mejores resultados que añadiendo además la información de SWN. Esto se debe al hecho de que, al contrario de *EmotiBlog*, el número de frases objetivas en *JRC* es mayor que el número de las subjetivas y, por lo tanto, la información que SWN suministra no tiene tanto impacto en el corpus. Mirando en todos los resultados en general (no sólo en los mejores) podemos observar los mismos aspectos que obtuvimos con *EmotiBlog* sobre las técnicas de *stemming*, *lematización* y RDS. También apreciamos una semejanza con los resultados de la Tabla 5 al evaluar las clasificaciones de forma combinada, pues en este caso también mejoran. En general, los resultados en estos experimentos son peores que en los realizados con *EmotiBlog*, a pesar de que el corpus de *JRC* es más grande. Esto es debido a que el proceso de anotación de *JRC* incluía unas reglas muy laxas para los anotadores que permitían mayores errores en el etiquetado. Estas reglas provocan que frases subjetivas puedan ser etiquetadas como objetivas, creando ruido en los modelos de AA. Esto no ocurre con *EmotiBlog* porque el proceso de anotación ha sido más cuidadoso y estricto. En la Tabla 7 resumimos los mejores resultados para cada corpus.

	EB Kyoto	EB Phones	EB Full	JRC
Objetividad	**0.6647**	0.6405	0.6274	0.6088
Polaridad	0.7602	**0.8093**	0.6374	0.5340
Obj+Pol	**0.5893**	0.5447	0.5493	0.5769

Tabla 7. Mejores resultados según clasificación y corpus.

5 *Conclusiones y Trabajos Futuros*

Es bien conocido que la importancia del AS se ha visto incrementada debido a la inmensa cantidad de información subjetiva disponible. Esto es debido a que es necesario explotar esta información en aplicaciones que trabajen con opiniones. Es por esto que en este artículo hemos evaluado las técnicas existentes enfocándonos en la detección y clasificación automática de frases con información subjetiva. Los corpus que hemos utilizado proceden de distintos dominios y géneros textuales y esto dificulta la tarea, especialmente en el dominio de los teléfonos móviles dónde tenemos expresiones de subjetividad más informales si los comparamos con el resto de corpus basados en comentarios de periódicos. Para entrenar y probar nuestro sistema de AA para la detección automática de datos subjetivos hemos utilizado el corpus *EmotiBlog Phones,* una extensión de *EmotiBlog*. Hemos procesado todas las combinaciones de TSR, tokenización y pesado de términos, llegando a un total de 1 millón de experimentos, pero por razones de espacio nos hemos limitado a presentar los resultados más relevantes. Como hemos demostrado, el AS es una tarea extremadamente compleja y hay muchas posibilidades de mejorar los resultados obtenidos. Con respecto a la mejora de la detección el objeto del discurso (target) nuestra intención es usar modelos de entrenamiento basados en secuencias de palabras (*n-gramas*, *modelos ocultos de Markov*, etc) para encontrar el tema principal del discurso de una opinión y poder hacer un estudio comparativo de las diferentes técnicas (que se usarán también para detectar los fenómenos lingüísticos basados secuencialidad, como las negaciones, la ironía y el sarcasmo).

Como trabajo futuro nuestra intención es unir los corpus (*EmotiBlog* y el *JRC*) además de otras colecciones disponibles para poder tener a disposición más datos para los modelos de AA y así obtener mayor precisión sobre los elementos que tienen en común en la anotación. Cabe mencionar que *EmotiBlog* contiene también textos extraídos de blogs en italiano y castellano, pero debido a la falta de recursos en estos idiomas y para hacer un estudio comparativo (*JRC* sólo contiene textos en inglés), hemos decidido utilizar únicamente la parte en inglés. En trabajos futuros tenemos la intención de explotar las colecciones en italiano y castellano de *EmotiBlog*, además de anotar más textos en nuevos idiomas. En este trabajo no hemos aprovechado la totalidad de etiquetas de granularidad fina que proporciona *EmotiBlog*. Por eso en futuras investigaciones nos proponemos tener en cuenta esta información para mejorar nuestros modelos de AA. Respecto a la fiabilidad de *EmotiBlog* para tareas de AS y su comparación con el corpus JRC, hemos observado que experimentando con las mismas técnicas en ambos corpus se han obtenido resultados muy parecidos. Este hecho nos muestra que el sistema de anotación y el proceso de etiquetado de *EmotiBlog* son válidos. En algunos casos los resultados obtenidos con *EmotiBlog* son mejores, confirmando nuestra hipótesis de que este corpus está basado en un sistema de anotación apropiado y sigue un sistema de anotación robusto. Esto nos anima a continuar nuestra investigación utilizando *EmotiBlog*.

6 *Referencias*

Agirre, E., Lopez de Lacalle, O., Fellbaum, C., Hsieh, S., Tesconi, M., Monachini, M., Vossen, P., Segers, R. 2010. SemEval-2010 Task 17: All-words Word Sense Disambiguation on a Specific Domain.

Abulaish, M., Jahiruddin, M., Doja, N. and Ahmad, T. 2009. Feature and Opinion Mining for Customer Review Summarization. PReMI 2009, LNCS 5909, pp. 219–224, 2009. Springer-Verlag Berlin Heidelberg.

Balahur A., and Montoyo A. 2008. Applying a Culture Dependent Emotion Triggers Database for Text Valence and Emotion Classification. In Proceedings of the AISB 2008 Symposium on Affective Language in Human and Machine, Aberdeen, Scotland.

Balahur A., Lloret E., Boldrini E., Montoyo A., Palomar M., Martínez-Barco P. 2009a. Summarizing Threads in Blogs Using Opinion Polarity. In Proceedings of ETTS workshop. RANLP.

Balahur, A., Boldrini, E., Montoyo, A., Martínez-Barco, P. 2009c. Opinion and Generic Question Answering systems: a performance analysis. In Proceedings of ACL, 2009, Singapore.

Balahur, A., Boldrini, E., Montoyo, A., Martínez-Barco, P. 2010b. Opinion Question Answering: Towards a Unified Approach. In Proceedings of the ECAI conference.

Balahur, A., Boldrini, E., Montoyo, A., Martínez-Barco 2009b. P. Cross-topic Opinion Mining for Realtime Human-Computer Interaction. ICEIS 2009.

Balahur Alexandra, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik van der Goot, Matina Halkia, Bruno Pouliquen & Jenya Belyaeva (2010c). Sentiment Analysis in the News. In: Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'2010), pp. 2216-2220. Valletta, Malta, 19-21 May 2010.

Boldrini, E., Balahur, A., Martínez-Barco, P., Montoyo, A. 2010. EmotiBlog: a finer-grained and more precise learning of subjectivity expression models. In Proceedings of LAW IV, ACL.

Boldrini, E., Balahur, A., Martínez-Barco, P., Montoyo, A. 2009a: EmotiBlog: an Annotation Scheme for Emotion Detection and Analysis in Non-traditional Textual Genres. In Proceedings of DMIN, Las Vegas.

Balahur, A., Boldrini, E., Montoyo, A., Martínez-Barco, P. 2010a. A Unified Proposal for Factoid and Opinionated Question Answering. In Proceedings of the COLING conference.

Boldrini E., Fernández J., Gómez J.M., Martínez-Barco P. 2009b. Machine Learning Techniques for Automatic Opinion Detection in Non-Traditional Textual Genres. In Proceedings of WOMSA 2009. Seville, Spain.

Chaovalit P, Zhou L. 2005. Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches. In Proceedings of HICSS-05.

Cui H., Mittal V., Datar M. 2006. Comparative Experiments on Sentiment Classification for Online Product Reviews. In Proceedings of the 21st National Conference on Artificial Intelligence AAAI.

Cerini S., Compagnoni V., Demontis A., Formentelli M., and Gandini G. 2007. Language resources and linguistic theory: Typology, second language acquisition. English linguistics (Forthcoming), chapter Micro-WNOp: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining. Franco Angeli Editore, Milano, IT.

Dave K., Lawrence S., Pennock, D. 2003. "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews". In Proceedings of WWW-03.

Esuli A., Sebastiani F. 2006. SentiWordNet: A Publicly Available Resource for Opinion Mining. In Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy.

Gamon M., Aue S., Corston-Oliver S., Ringger E. 2005. Mining Customer Opinions from Free Text. Lecture Notes in Computer Science.

Galavotti, L., Sebastiani, F., and Simi, M. 2000. Experiments on the use of feature selection and negative evidence in automated text categorization. In Proceedings of ECDL-00, 4th European Conference on Research and Advanced Technology for Digital Libraries (Lisbon, Portugal, 2000), 59–68.

Gómez, J.M.; Buscaldi, Bisbal, E.; D.; Rosso P.; Sanchis E. QUASAR: The Question Answering System of the Universidad Politécnica de Va-

lencia. In Accessing Multilingual Information Repositories. LNCS 2006. 439-448.

Hatzivassiloglou V., Wiebe J. 2000. Effects of adjective orientation and gradability on sentence subjectivity. In Proceedings of COLING.

Liu 2006. Web Data Mining book. Chapter 11

Liu, B. (2007). Web Data Mining. Exploring Hyperlinks, Contents and Usage Data. Springer, first edition.

Miller, G.A. 1995. WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41

Mladenic, D. 1998. Feature subset selection in text learning. In Proceedings of ECML-98, 10th European Conference on Machine Learning (Chemnitz, Germany, 1998), 95–100.

Mullen T., Collier N. 2004. Sentiment Analysis Using Support Vector Machines with Diverse Information Sources. In Proceedings of EMNLP.

Ng V., Dasgupta S. and Arifin S. M. 2006. Examining the Role of Linguistics Knowledge Sources in the Automatic Identification and Classification of Reviews. In the proceedings of the ACL, Sydney.

Ohana, B., Tierney, B. 2009. Sentiment classification of reviews using SentiWordNet, T&T Conference, Dublin Institute of Technology, Dublin, Ireland, 22nd.-23rd.

Pang B and Lee L. 2003 Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In Proceedings of the 43rd Annual Meeting of the ACL, pages 115–124.

Pang B., Lee L, Vaithyanathan, S. 2002. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of EMNLP-02, the Conference on Empirical Methods in Natural Language Processing.

Qiu, G., Liu, B., Bu, J., Chen, C. 2006. Opinion Word Expansion and Target Extraction through Double Propagation. Association for Computational Linguistics

Riloff E. and Wiebe J. 2003. Learning Extraction Patterns for Subjective Expressions. In Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing.

Salton, G. and Buckley, C. (1988). "Term Weighting Approaches in Automatic Text Retrieval." In: Information Processing and Management, 24(5).

Strapparava C. Valitutti A. 2004. WordNet-Affect: an affective extension of WordNet. In Proceedings ofthe 4th International Conference on Language Resources and Evaluation, LREC.

Turney P. 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. ACL 2002: 417-424.

Wiebe, J. (1994). Tracking point of view in narrative. Computational Linguistics 20 (2): 233-287.

Wilson, T., Wiebe, J., Hwa, R. 2006. Recognizing strong and weak opinion clauses. Computational Intelligence 22 (2): 73-99.

Yang, J. and and Pedersen, O. 1997. A comparative study of feature selection in text categorization. In: ICML ’97: Proceedings of the Fourteenth International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 412–420.