Evaluating the Robustness of EmotiBlog for Sentiment Analysis and Opinion Mining

Ester Boldrini, Javi Fernández, José Manuel Gómez and Patricio Martínez-Barco
GPLSI – University of Alicante
{eboldrini; javifm; jmgomez; patricio}@dlsi.ua.es

Abstract

Preliminary research demonstrated the EmotiBlog annotated corpus relevance as a Machine Learning resource to detect subjective data. In this paper we compare EmotiBlog with the JRC Quotes corpus in order to check the robustness of its annotation. We concentrate on its coarse-grained labels and carry out a deep Machine Learning experimentation also with the inclusion of lexical resources. The results obtained show a similarity with the ones obtained with the JRC Quotes corpus demonstrating the EmotiBlog validity as a resource for the SA task.

1 Introduction and Motivation

Due to the birth of the Web 2.0 and the wide employment of the new textual genres we have an exponential increase of the subjective information. We also have a recent explosion of interest in Sentiment Analysis (SA), a subtask of Natural Language Processing (NLP), in charge of identifying the opinions related to a specific target (Liu, 2006). Subjective data has a great potential; it can be exploited by business organizations or individuals, for ads placements, but also for the Opinion Retrieval/Search, etc (Liu, 2007). Our research is motivated by the lack of resources, methods and tools to effectively process subjective information. Our main purpose is to demonstrate that the *EmotiBlog* corpus can be a robust resource to overcome the challenges SA brings. For these first experiments we take into account its coarse-grained annotation; however in the future we will concentrate on the finer-grained annotation. We train our Machine Learning (ML) system with *EmotiBlog Kyoto*[1] and *EmotiBlog Phones*[2] corpora, but also with the *JRC Quotes*[3] collection. These experiments are possible since the corpora share some common annotated elements (Section 3), thus allowing a larger dataset and comparable results. Then, we train our system with some of the features of *EmotiBlog* and we also integrate 2 lexical resources to reach a wider coverage. We also employ NLP techniques (stemmer, lemmatiser, bag of words, etc.) to improve the results obtained with the supervised ML models. In previous works it has been demonstrated that *EmotiBlog* is a beneficial resource for Opinionated Question Answering (OQA) as stated Balahur et al. (2009c and 2010) or Automatic Opinionated Summarization (Balahur et al. 2009a). Thus, our first objective is to demonstrate that *EmotiBlog* is a useful resource to train ML systems for SA. The combination of training from *EmotiBlog* and *JRC Quotes* is beneficial since it provides more data for the common labelled elements. As a consequence, our second purpose is to demonstrate that a deeper text classification is crucial (Section 2). We believe there is a need for determining the emotion intensity (*high/medium/ low*) and the emotion type apart from other elements presented in Boldrini et al. (2010).

2 Related Work

The first step of SA research consists in building up lexical resources of affect, such as *WordNet Affect* (Strapparava and Valitutti, 2004), *SentiWordNet* (Esuli and Sebastiani, 2006), or *Micro-WNOP* (Cerini et. al., 2007). Moreover, (Wiebe 2004) focused the idea of subjectivity around that of private states setting the benchmark for subjectivity analysis. Authors show that the discrimination between objective/subjective discourses is crucial for the SA, as part of Opinion Information Retrieval (TREC Blog tracks[4] and the TAC 2008 competitions[5]), Information

[1] The *EmotiBlog* corpus is composed by blog posts on the Kyoto Protocol, Elections in Zimbabwe and USA election, but for this research we only use the *EmotiBlog* Kyoto (about the Kyoto Protocol)
[2] it is an EmotiBlog extension with reviews of mobiles

[3] http://langtech.jrc.ec.europa.eu/JRC_Resources.html
[4] http://trec.nist.gov/data/blog.html
[5] http://www.nist.gov/tac/

Proceedings of Recent Advances in Natural Language Processing, pages 521–526,
Hissar, Bulgaria, 12-14 September 2011.

Extraction (Riloff and Wiebe, 2003) and QA (Stoyanov et al., 2005) systems. Related work also includes sentiment classification using unsupervised methods (Turney, 2002), ML techniques (Pang and Lee, 2002), scoring of features (Dave, Lawrence and Pennock, 2003), using PMI, or syntactic relations and other attributes with SVM (Mullen and Collier, 2004). Research in classification at a document level included sentiment classification of reviews (Ng, Dasgupta and Arifin, 2006). Neviarouskaya (2010) classified texts using fine-grained attitude labels basing its work on the compositionality principle and an approach based on the rules elaborated for semantically distinct verb classes and Tokuhisa (2008) proposed a data-oriented method for inferring the emotion of a speaker conversing with a dialogue system from the semantic content of an utterance. Wilson et al 2009 worked on mixed results and for Ghazi et al 2010 the hierarchy was better on two datasets. Our work starts from the conclusions drawn by (Boldrini et al 2010). They showed that the different levels of annotation that *EmotiBlog* contains offers important information on the structure of subjective texts, leading to an improvement of the performance of systems trained on it.

3 Corpora

The corpus we mainly employed in this research is *EmotiBlog*[6] *Kyoto* extended with the collection of mobile phones (*EmotiBlog Phones*): the *EmotiBlog Full*. The first part is a collection of blog posts in English extracted from the web containing opinions about the Kyoto Protocol, while the second part is composed by reviews of mobiles phones extracted from Amazon[7]. *EmotiBlog* annotation model contemplates *document/sentence/ element levels of annotation* (Boldrini et al. 2010), and distinguishes *objective/subjective* discourse Boldrini et al. (2009a). For all of these elements, common attributes are annotated: *polarity*, *degree* and *emotion*. Two experienced annotators labelled this collection and previous work done by Boldrini et al, 2009a) detected a high percentage of inter-annotator agreement, thus proving a reliable tagging. We also used the *JRC* Quotes corpus[8] (1590 English quotations extracted from the news and manually annotated for the sentiment expressed towards entities mentioned inside the quotation) (Balahur et al., 2010c).

[6] Available on request from authors

[7] www.amazon.com

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

4 ML Experiments and Discussion

For demonstrating that *EmotiBlog* is a robust resource for ML, we performed a series of experiments using different approaches, corpus elements and resources.

4.1 EmotiBlog without Semantic Information

First we used *EmotiBlog Kyoto* and *Phones* and a combination of them (*EmotiBlog Full*).

	Classification	Samples	Categories
EmotiBlog Kyoto	Objectivity	557	2
	Polarity	203	2
	Degree	209	3
	Emotion	132	5
	Obj+Pol	550	3
	Obj+Pol+Deg	549	6
EmotiBlog Phones	Objectivity	418	2
	Polarity	245	2
	Degree	236	3
	Emotion	234	4
	Obj+Pol	417	3
	Obj+Pol+Deg	409	7
EmotiBlog Full	Objectivity	974	2
	Polarity	448	2
	Degree	445	3
	Emotion	366	5
	Obj+Pol	967	3
	Obj+Pol+Deg	958	7

Table 1: # of samples and categories by classification

Classifying either objectivity or polarity is simpler than degree or emotion due to the smaller number of categories these last ones contain. For the polarity evaluation we need the objectivity to have been evaluated previously (*subjective/objective* discrimination) to work with the selected subjective sentences. The same situation applies for the *degree*, since we have to determine if it refers to the *positive*/*negative* polarity. The consequence of this process is that the classification errors of polarity and objectivity are propagated affecting the final degree evaluation. Thus we combined the classifications to check if this approach improves the results for evaluating *polarity* and *degree*. We combined *polarity* with *objectivity* (*Obj+Pol*), with 3 resulting categories: *objective*, *positive* and *negative*. We also combined *degree+objectivity+pola-rity* with the 7 resulting categories.

In this first step we use the classic *bag of words* (**word**) and to reduce the dimensionality we employ *stemming* (**stem**), *lemmatization* (**lemma**) and *dimensionality reduction by term selection* (TSR) methods. For TSR, we compare two approaches, *Information Gain* (**ig**) and *Chi Square* (**x2**), since they reduce the dimensionality substantially with no loss of effectiveness (Yang and Pedersen, 1997). We have applied these techniques with a different number of selected terms for each of them (**ig50**, **ig100**, … **ig1000**). For weighting these features we evaluate the most common methods: *binary weighting* (**binary**), *tf/idf* (**tfidf**) and *tf/idf normalized* (**tfidfn**) (Salton and Buckley, 1988). We also included as weighting technique the one use by Gómez et al. (2006) in IR tasks to evaluate its reliability in different domains (**jirs**). It is similar to *tf/idf* but it does not take into account term frequencies. We will also use its normalized version (**jirsn**). As supervised learning method we use *Support Vector Machines* (SVM) due to its good performance in text categorization (Sebastiani, 2002) and the promising results obtained in previous studies (Boldrini et al. 2009b). The best results are shown in in Table 2. Due to the high number of experiments (about 1 million) and ML adjustment parameters carried out, for space reasons we present only the best performance obtained. As baseline we employed a classifier that always chooses the most frequent class. Our best results are obtained with *lemmatisation* (high number of features) and *stemming* (with few features). Experiments with TSR obtain higher scores, without any significant difference between *x2* and *ig*. The number of features selected by TSR range s between 100 and 800, depending on the number of classes and samples of the classification (the bigger they are, the more features are needed). In addition, if we do not apply *stemmer* or *lemmatiser*, the number of features must be increased for better results. Using TSR improves the results. The *tf/idf* performs better except for the polarity, where *tf/idf normalised* works better. No significant differences were found between using the normalised version of *tf/idf*, *jirs* or *jirs normalised*. In general any feature weight technique works better than the *binary* one, giving similar results independently from the method selected. We can observe that the results obtained with *Kyoto* and *Phones* corpora separately are better than using both corpora (*Full*) to build the ML model. Moreover, the learned ML models of *Kyoto* and *Phones* corpora are more specialized. They are only appropriate for classifying opinions about their own domain, the Kyoto. As we can deduce from the experiments, objectivity and polarity classifications evaluation is less problematic due to the low number of categories of each one of them. In addition, once we have detected the objectivity, the polarity is easier to determinate although the number of samples for polarity is a 41% smaller and both have the same number of categories. The first task is more complex, because the feature space vectors in the two objectivity categories are closer and we have more ambiguity in objectivity classification than in polarity classification. Terms as '*bad*', '*good*', '*excellent*' or '*awful*' clearly determine the polarity of the sentences but it is more difficult to find this kind of terms for the objectivity. Although the combinations of categories (*Obj+Pol* and *Obj+Pol+Deg*) give lower *f-measure*, this does not mean that these approaches are not adequate. In order to obtain the score for polarity and degree in Table 2, we

	Classification	Baseline	word		lemma		stem	
		f-measure	f-measure	techniques	f-measure	techniques	f-measure	techniques
EmotiBlog Kyoto	Objectivity	0.4783	0.6440	tfidf, chi950	0.6425	tfidfn	**0.6577**	**tfidfn, chi250**
	Polarity	0.5694	0.7116	jirsn, ig400	0.6942	tfidf, ig200	**0.7197**	**tfidf, ig500**
	Degree	0.3413	0.5884	tfidf, ig900	**0.6296**	**tfidf, ig350**	0.6146	tfidfn, ig600
	Emotion	0.1480	0.4437	tfidfn, ig350	**0.4665**	**jirsn, ig650**	0.4520	jirsn, ig650
	Obj+Pol	0.4881	0.5914	jirsn, ig600	0.5899	tfidfn, ig750	**0.6064**	**jirsn, ig250**
	Obj+Pol+Deg	0.4896	0.5612	jirsn	**0.5626**	**jirsn**	0.5433	tfidf, ig700
EmotiBlog Phones	Objectivity	0.4361	0.6200	jirsn, ig900	**0.6405**	**tfidfn, chi500**	0.6368	tfidfn, ig600
	Polarity	0.7224	**0.7746**	**tfidf, ig250**	0.7719	tfidfn	0.7516	tfidfn, ig500
	Degree	0.5153	0.6156	tfidfn	**0.6174**	**jirsn, ig650**	0.6150	tfidf, ig650
	Emotion	0.7337	0.7555	jirsn, ig450	**0.7828**	**jirsn, ig150**	0.7535	tfidf, ig350
	Obj+Pol	0.3057	0.5287	tfidf, ig650	**0.5344**	**tfidfn, ig900**	0.5227	tfidf, ig850
	Obj+Pol+Deg	0.2490	0.4395	tfidf, ig700	0.4424	tfidf	**0.4557**	**tfidf, ig600**
EmotiBlog Full	Objectivity	0.3705	0.5964	jirsn, ig150	0.6080	jirsn, chi100	**0.6229**	**jirsn, ig350**
	Polarity	0.3880	0.6109	tfidfn, ig1000	**0.6196**	**tfidf, chi100**	0.6138	tfidf, chi50
	Degree	0.4310	0.5655	jirsn	0.5526	jirsn	**0.5775**	**jirsn, ig450**
	Emotion	0.3990	0.5675	jirsn, ig850	**0.5712**	**tfidfn, ig800**	0.5644	jirsn, ig800
	Obj+Pol	0.3749	0.5332	tfidf	0.5381	tfidf, ig700	**0.5431**	**tfidf**
	Obj+Pol+Deg	0.3807	0.4794	tfidf, ig700	0.4903	tfidf	**0.4923**	**jirsn**

Table 2: Experiments without semantic information

preselected only the subjective sentences for the polarity and degree evaluation, not possible in the real-world. We would need first to automatically classify the objectivity, then the polarity and the degree. This methodology drags errors in each evaluation. If we calculate the *precision* (P) instead of the *f-measure* of the best experiment for each category separately and obtain their final precision by propagating the error multiplying their precisions, the polarity measure does not seem to be so good. It is important to underline that, for the propagation of the objectivity categories, we only take into account the subjective precision and not the objective one (when we evaluate objectivity and polarity using the *Full* corpus we obtain a precision of **0.71** and **0.72** respectively). Therefore, the propagated precision would be the product of these values (0.51), which is 12% lower than evaluating *Obj+Pol* together (0.58). This is more significant if we evaluate degree separately, which gives us a precision 37% lower.

	Combination	Precision
EB Kyoto	P(Obj) · P(Pol)	0.4352
	P(Obj+Pol)	**0.6113**
	P(Obj) · P(Pol) · P(Deg)	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(Deg)	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(Deg)	0.3097
	P(Obj+Pol+Deg)	**0.4912**

Table 3: Precisions by combination of categories

In Table 3 we show the best results with the 3 main corpora. These improvements appear in all evaluations independently from the corpus and techniques used. The combination of categories improves the final results from 8.34% to 68.39%. The more categories are combined the bigger is the improvement because in the case of separate categories, the ML process has no information about the rest of categories when is learning for only one of them. When combining several categories we are adding this valuable information to the ML process and removing an important part of the propagation error.

4.2 EmotiBlog with Semantic Information

In order to check the impact of including the semantic relation as learning feature, we group features by their semantic relations, to increase the coverage and reduce the samples' dimensionality. The challenge here is *Word Sense Disambiguation* (WSD). We suppose that choosing the wrong sense of a term would introduce noise in the evaluation and a lower performance. But if we include all term senses term in the set of features, the TSR could remove the not useful ones (this disambiguation method would be adequate). We used two lexical resources: *WordNet* (WN) and *SentiWordNet* (SWN). The first one since it contains a huge quantity of semantic relations between English terms, and the second since the use of this specific OM resource demonstrated to improve the results of OM systems (Abulaish et al. 2009). It assigns to some of the synsets of WN three sentiment scores: *positivity*, *negativity* and *objectivity*. As the synsets in SWN are only the opinionated ones, we want to test if expanding only with those ones can improve the results. In addition, we want to introduce the sentiment scores into the ML system by adding them as new attributes. For example, if we get a synset *S* with a positivity score of 0.25 and a negativity score of 0.75, we add a feature called *S* (with the score given by the weighting technique) but also two more features: *S-negative* and *S-positive* with their negative and positive scores respectively. These experiments with lexical resources have been carried out with five different configurations using: only SWN synsets (**swn**), only WN synsets (**wn**), both SWN and WN synsets (**swn+wn**), only SWN synsets including sentiment scores (**swn+scores**) and both SWN and WN synsets including also the mentioned sentiment scores (**swn+wn+scores**). In case a term is not found in any of the lexical resources, then its lemma is used. Moreover, to solve the ambiguity, two techniques have been adopted: including all its senses and let the TSR methods perform the disambiguation (mentioned **swn**, **wn**, **swn+wn**, **swn+scores** and **swn+wn+scores**), but also including only the most frequent sense for each term (**swn1**, **wn1**, **swn1+wn1**, **swn1+scores** and **swn1+wn1+scores**).

Except for a few cases, the semantic information from WN and SWN improves the final results (+7.12%). We observed that the experiments using semantic information are always in the top results. Using only WN does not perform as well as with SWN, because it only contains information about subjective features, an important thing when selecting the best features for the classification task. From Table 4 we notice that TSR is present in almost all experiments with semantic information. Thus TSR techniques are adequate approximations for removing noise from the training corpus features. Again, the weighting techniques do not seem to have a big influence in opinion classification, but *tf/idf* and *jirs* perform always better than the *binary* approach. The best results include the lexical resources (always in the top positions). In Table 4 we see that SWN is present in all the best results, and the sentiment scores in 55% of them. Moreover, SWN and its scores appear in almost all best results for *EmotiBlog Full*. This technique seems to be better for not domain-specific corpus. It is important to stress upon the fact that methods, which use *ig* and *x2* improve the majority of the results confirming our hypothesis they are adequate for disambiguation.

	Classification	f-measure	Techniques
EmotiBlog Kyoto	Objectivity	0.6647	swn+wn+scores, tfidf, chi900
	Polarity	0.7602	swn1, tfidfn, chi550
	Degree	0.6609	swn1, jirsn, ig550
	Emotion	0.4997	swn, tfidf, chi450
	Obj+Pol	0.5893	swn, tfidfn
	Obj+Pol+Deg	0.5488	swn1+wn1, tfidf
EmotiBlog Phones	Objectivity	0.6405	swn1+wn1+scores, jirsn, ig1000
	Polarity	0.8093	swn+scores, tfidfn, ig550
	Degree	0.6306	swn1+wn1, tfidfn, ig150
	Emotion	0.8133	swn+wn+scores, jirsn, ig350
	Obj+Pol	0.5447	swn+wn+scores, tfidfn, chi200
	Obj+Pol+Deg	0.4445	swn1, jirsn
EmotiBlog Full	Objectivity	0.6274	swn+wn, jirsn, chi650
	Polarity	0.6374	swn1+scores, jirsn, chi350
	Degree	0.6101	swn1+wn1+scores, tfidf, ig1000
	Emotion	0.5747	swn+wn+scores, jirsn, ig450
	Obj+Pol	0.5493	swn+wn+scores, tfidf, chi950
	Obj+Pol+Deg	0.4980	swn+wn+scores, jirsn

Table 4: Results with semantic information

4.3 Experiments with the JRC Corpus

We have applied the same ML techniques with the *JRC Quotes* corpus. We can observe in first instance that experiments adding lexical resources, either WN or SWN, obtain better score than experiments without it (Table 5). Using only WN performs better than adding SWN (because the number of objective sentences in *JRC Quotes* is greater than the number of subjective ones). That is why the information that SWN provides does not have the same impact with this corpus. The *binary* weighting technique also performs worse than the rest of techniques, which seem to be indifferent for *EmotiBlog*. The precisions combining the classifications objectivity and polarity are also better than calculating the precisions separately and propagating the errors. In general, the *f-measure* is worse than in the ones with *EmotiBlog* despite the fact that the *JRC Quotes* is bigger.

	Classification	f-measure	Techniques
Baseline	Objectivity	0.5363	-
	Polarity	0.3880	-
	Obj+Pol	0.5363	-
Word	Objectivity	0.6022	tfidfn, ig950
	Polarity	0.5163	jirsn
	Obj+Pol	0.5648	tfidfn, ig100
Lemma	Objectivity	0.6049	jirsn
	Polarity	0.5240	tdidfn, ig800
	Obj+Pol	0.5697	jirs
Stem	Objectivity	0.6066	jirsn
	Polarity	0.5236	tfidfn, ig450
	Obj+Pol	0.5672	tfidf
WN	Objectivity	**0.6088**	wn1, jirsn, ig650
	Polarity	**0.5340**	wn1, tfidfn, ig800
	Obj+Pol	**0.5769**	wn1, jirsn, ig700
SWN + WN	Objectivity	0.6054	swn1+wn1, jirsn
	Polarity	0.5258	swn+wn+scores, jirsn
	Obj+Pol	0.5726	swn1+scores, jirsn

Table 5: Experiments with JRC

The cause of this is that its annotation process instructions are: *If the annotator doubts when deciding if a sentence is objective or subjective, then he must leave it blank, and If a sentence has been left blank, then the sentence is supposed to be objective.* These rules cause several subjective sentences to be tagged as objective creating noise to our ML approaches.

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

Table 6. Comparison of best results per classification/corpus.

5 Conclusions and Future Works

The corpora we employed are *EmotiBlog* and the *JRC Quotes* collection. We processed all the combinations of TSR, tokenisation and term weighting for a total of 1M experiments, showing only the most significant results. The SA is a challenging task and there is room for improvement. For target detection we will employ learning models based on sequence of words (*n-grams*, *Hidden Markov Models*, etc.) to find the topic of published opinion and making a comparative assessment of different techniques. We will also merge both corpora (*EmotiBlog* and *JRC Quotes*) and other collections to have more data for the ML models. We will take into account the totality of the *EmotiBlog* annotation to improve our ML models with this fine-grained data. We observed that experimenting with the same techniques both of the corpora obtained close or higher results demonstrating that the *EmotiBlog* is a valid resource.

References

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., 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.

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

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.

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.

Gómez, J.M.; Buscaldi, Bisbal, E.; D.; Rosso P.; Sanchis E. QUASAR: The Question Answering System of the Universidad Politécnica de Valencia. In Accessing Multilingual Information Repositories. LNCS 2006. 439-448.

Liu 2006. Web Data Mining book. Chapter 11

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

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

Neviarouskaya, A., Prendinger, H. and Ishizuka, M. 2010. User study on AffectIM, an avatar-based Instant Messaging system employing rule-based affect sensing from text. Int. Journal of Human-Computer Studies 68(7):432–450.

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.

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.

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.

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.