EmotiBlog: Towards a Finer-Grained Sentiment Analysis and its Application to Opinion Mining

1 Introduction and Motivation

The exponential growth of the subjective information freely available on the Web and the employment of new textual genres has created an explosion of interest in Sentiment Analysis (SA). This is a task 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 properly threat subjective data. Our main purpose is to demonstrate that EmotiBlog - a corpus annotated with the EmotiBlog annotation schema **for detecting** subjectivity in the new textual genres- can be successfully employed to overcome the challenges of fine-grained SA. We also want to demonstrate that it contributes to solve the shortage of coarse-grained annotation data and improves the Opinion Mining (OM) task. In order to achieve this, we train our Machine Learning system with *EmotiBlog Kyoto*[1] and *EmotiBlog Phones,* but also with the *JRC*[2] corpus[3]. Then, we train with the *EmotiBlog corpus* finer-grained features (not available in the *JRC* annotation) and we integrate *SentiWN* (Esuli and Sebastiani, 2006) and *WordNet* (Miller, 1995). We also employ NLP techniques (stemmer, lemmatiser, bag of words, etc.) to improve the results obtained with the supervised ML models. After that, we apply the trained system to the OM task (using a collection of reviews from *Amazon*[4]), to automatically detect the users' points of view about a mobile phone or one/more of its features. In previous works it has been showed that *EmotiBlog* is a beneficial resource for Opinionated Question Answering (OQA), as stated Balahur et al. (2009 c and 2010a,b) and for Automatic Summarization of subjective content (Balahur et al. 2009a). Thus, the first objective of our research is to demonstrate that *EmotiBlog* is a useful resource to train ML systems for OM applications. Most work done in OM only concentrated on classifying polarity of sentiments into *positive/negative*, thus our second objective is to demonstrate that the combination of training (*EmotiBlog* and *JRC)* is beneficial since we have more data for the common elements, but also a finer-grained analysis, assured by *EmotiBlog*. As a consequence, our third purpose is to demonstrate that a deeper text classification for the OM task is essential. There is the need for *positive/negative* text categories, but also *emotion intensity* (*high/medium/low)*, the *emotion type* (Boldrini et al, 2009a) and the annotation of the linguistic elements that give the subjectivity to the discourse. The complete list of elements is presented in Boldrini et al. (2010). Finally the fourth objective of this research is the implementation of an OM application prototype (which will reinforce the system utility) for retrieving opinions on a product or its features continuing the work proposed by Balahur et al. (2009b).

2 Corpora

The corpus (in English) we mainly employed in this research is *EmotiBlog Kyoto* extended with the collection of mobile phones reviews extracted from *Amazon* (*EmotiBlog Phones*)[5]. It allows the annotation at *document/sentence/element level* (Boldrini et al. 2010), distinguishing between *objective/subjective* discourse. A list of tags for the subjective elements is available

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

[3] feasible since they have common tags and this will allow us to have a larger data set for the common annotated elements

[4] www.amazon.com

[5] Available on request from authors

(Boldrini et al, 2009a) – *source*, *topic*, *verbs*, *nouns*, *adjectives*, *adverbs*, *sayings*, *collocations*, etc. For all of these elements, common attributes are annotated: *polarity*, *degree* & *emotion*. Table 1 presents the size of the corpus and its subjective elements.

Table 1: corpus overview

EB Kyoto	**Ws**: 12328 **Sub**: 210 **Ps**: 62 **Ng**: 141 **Obj**: 347 **Ph**: 692					
	Adj	**Noun**	**Adv**	**Prep**	**Pron**	**Verb**
	161	154	70	13	52	140
EB Phones	**Ws**: 7759 **Sub**: 246 **Ps**: 198 **Ng**: 47 **Obj**: 172 **Ph**: 521					
	Adj	**Noun**	**Adv**	**Prep**	**Pron**	**Verb**
	212	61	94	0	0	39
EB Full	**Ws**: 20087 **Sub**: 455 **Ps**: 260 **Ng**: 188 **Obj**: 519 **Ph**: 1213					
	Adj	**Noun**	**Adv**	**Prep**	**Pron**	**Verb**
	373	215	164	13	52	179
JRC	**Ws**: 39214 **Sub**: 427 **Ps**: 193 **Ng**: 234 **Obj**: 863 **Ph**: 427					
	Adj	**Noun**	**Adv**	**Prep**	**Pron**	**Verb**
	0	0	0	0	0	0

Where *Ws*, *Sub*, *Ps*, *Ng*, *Obj*, *Ph*, *Adj*, *Noun*, *Adv*, *Prep*, *Pron* and *Verb* correspond to the number of words, subjective/positive/negative/objective sentences, total of phrases, adjectives, nouns, adverbs, prepositions, pronouns and verbs which have been annotated. We also used the *JRC* quotes[6], a set of 1590 English language quotations extracted automatically from the news and manually annotated for the sentiment expressed towards entities mentioned inside the quotation. The *JRC* is labelled in a coarse-grained way –if compared with *EmotiBlog*- thus, we use it to train our ML system for the element it has in common with *EmotiBlog,* and we then improve the training adding the *EmotiBlog* finer-grained elements.

3 ML Experiments

In order to demonstrate that *EmotiBlog* is valuable resource for ML, we perform a series of experiments with different approaches, corpus elements and resources. First, we employ the bag of word extracted from the train corpus (*EmotiBlog*) and use basic techniques: tokenisation and dimensionality reduction by term selection (TSR) methods. Table 2 shows the most significant results. We used Support Vector Machine (SVM) due to the promising results obtained by Boldrini et al. (2009b). For TSR, we compared Information Gain (IG) and Chi Square (X2) for reducing the dimensionality substantially with no loss of effectiveness (Yang and Pedersen, 1997). For the feature weight needed by SVM we adopted the binary weight, assigning 1 to the feature that appears in the sample and 0 otherwise; tf/idf, which sets the tf/idf value (Salton and Buckley, 1988) of each feature if it appears in the sample and 0 otherwise. For tf/idf approach, we have also used the normalized one, tf/idfn (Sebastiani, 2002).

Table 2: ML experiments results

EM elements	**F-measure**	**Precision**	**Recall**	**Classes**
objectivity	0.6223	0.6601	0.642	**2**
polarity	0.6196	0.7209	0.6612	**2**
degree	0.5709	0.5985	0.6026	**3**
emotion	0.5712	0.6096	0.6433	**3**
obj+pol	0.5431	0.5771	0.5866	**3**
obj+pol+deg	0.4922	0.5018	0.5612	**9**

Table 2 shows the best results obtained using lemmatiser or stemmer. The stemmer improves the results in evaluation with few features and the lemmatiser when features are reduced. The tf/idf performs better in each evaluation, except for the polarity where td/idf normalised set is used. TSR systems obtain high results in each case without any significant differences between X2 and IG and the range of featured has changed between 100 and 800 depending on the number of classes. From the results in the mix of elements (*objectivity*/*polarity*) or *objectivity*/*polarity*/*degree* we can deduce that learning a model, which combines such elements improves the performance. To evaluate the degree we will first determine if the sentence is

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

subjective/objective, its *polarity* and *intensity*, thus increasing the possibility of mistakes. In order to check the impact of including the semantic relation as learning features, we believe that, grouping features by their semantic relations will increase the coverage in the test corpus a part from reducing the samples' dimensionality. The challenge at this point is Word Sense Disambiguation (WSD) due the poor results that these systems traditionally obtain in international competitions (Agirre et al. 2010). Choosing the wrong sense of a term would introduce noise in the evaluation and thus a low performance. The question is that if we include all senses of a term in the set of features, if the TSR would choose the correct ones. If we use all *WordNet* senses of each term as learning features, then the TSR methods IG/X2 could remove the not useful senses to classify the sample in the correct class. In this case this disambiguation methods would be adequate. The evaluation summarized in Table 3 focuses on solving these questions.

Table 3: Results with lexical resources

EM elements	F1	Precision	Recall	Resources
objectivity	0.6261	0.6538	0.6409	**swn+wn**
polarity	0.6195	0.6809	0.6481	**swn+wn**
degree	0.6101	0.6287	0.6381	**swn1+wn1**
emotion	0.5637	0.6114	0.6239	**swn+wn**
obj+pol	0.5493	0.5946	0.5959	**swn+wn**
obj+pol+deg	0.4802	0.4724	0.5458	**swn+wn**

We used two lexical resources: *WordNet* and *SentiWordNet*. The first one because it contains a huge quantity of semantic relations between English terms; and the second one since, the use of this specific OM resource demonstrated to improve the results of OM systems. *SentiWordNet* also assigns to each synset of WordNet 3 sentiment scores: *positivity/negativity/objectivity*. As we can observe in Table 3, experiments have been carried out with 5 different configurations using: i) *SentiWN* synsets, ii) WN synsets, iii) a combination of both, iv) *SentiWN* synsets+scores and v) *SentiWN* synset+scores combined with *WN* synsets. In configuration i) if the lemma of a word appeared in *SentiWN*, it was replaced by the related synset. If a word is not found the lemma it will be left. For the experiment ii) we use WN instead of *SentiWN* and repeat the previous process. In the next experiment (iii), we first compare terms with *SentiWN* and, only if, terms are not found, *WN* is used. In the fourth case, if a term appeared in *SentiWN* it was replaced by the related synset and their associated polarity scores as new attributes. Finally, in the last configuration (iv), we applied the previous process but, if a term does not appear in *SentiWN*, we checked if it does in *WN* and if found, it was replaced only by its synset. As always, in case the word was not found, its lemma was left. In order to solve the ambiguity, 3 techniques have been adopted: taking into account only the most frequently sense, including all senses, or including all senses but using both TSR techniques (IG & X2) with the goal of checking these methods as disambiguators. As we can see in previous table 3, most of experiment using *SentiWN* and *WN* improve slightly the results if compared to Table 2. Methods, which use IG and X2 improve the majority of the results confirming our hypothesis they are adequate for disambiguation. Finally, we have applied these models with the *JRC* corpus. These experiments obtain **0.70** and **0.66** of f-measure for *objectivity* and *polarity* respectively. Although the results with *JRC* are slightly higher than the ones with *EmotiBlog*, this is because *EmotiBlog* has a finer-grained text analysis and is much smaller than the *JRC*. Moreover, *JRC* is based on more formal texts, which do not have the language variability that *EmotiBlog* has. In the future we hope to improve the results increasing the *EmotiBlog* corpus with more samples and domains.

4 GPLSI EmotiReview

After having performed the previous experiments, we created an on-line application (*GPLSI EmotiReview*[7]) for exploiting the learnt models to the real life –adapting it with a domain onthology. *GPLSI EmotiReview* is the first version of a prototype of an OM system that could be employed to extract the overall opinion or some features of mobile phone. The system is divided into 2 modules: i)the intelligent crawler tracks user's opinions in specialised Web pages

[7] http://intime.dlsi.ua.es:8080/emotireview

(offline); and ii) the users queries are processed and the requested opinions given back the user (real-time). The crawler includes ML tools to detect which comments are subjective, discriminates them into *positive/negative/level of subjectivity* of the emotion expressed. In order to detect the emotion target, we follow the approach by Qiu et al. (2006) who use *Minipar*[8] to detect the syntactic relation between terms. In our case, thanks to *EmotiBlog* we have the subjective words annotated and we use Minipar to find their syntactic relations. In order to improve this process we link the subjective terms with an ontology we manually created (which includes all the features of mobiles) and in this way we understand better which adjective is related with which feature of mobiles. If we cannot find any relation, the target will be the general one of the document. If some feature product or one of its feature (screen, battery, memory) is detected inside the window near to a subjective expression, this expression will be about this target and if not it will refer to the product in general. In order to obtain the relation of a product and its features a specific ontology has been built (about the smart phones domain) which will also be extended in the future to be also useful for other areas. Once the information is collected, the system uses Lucene[9] (Hatcher and Gospodnetic, 2004) as search engine to find (between the stored products), to retrieve the products (similar to the query) and give back the related opinions as well as the general evaluation and its specific evaluations, if applicable.

5 Conclusions and Future Works

The first contribution this paper brings is the employment of *EmotiBlog* –a collection of blog posts labelled with the homonymous annotation schema- and the *JRC* corpus. They have been employed to train and test our ML system for the automatic detection of subjective data in the *EmotiBlogPhones* corpus, an extension of the *EmotiBlog*. We used both corpora to train the system regarding their common labelled elements and then *EmotiBlog* for a finer-grained text analysis, since it contains a finer-grained annotation. We processed all the combinations of TSR, tokenisation and term weight for a total of 660000 experiments, but due to space reasons we showed the most significant. Another contribution is the implementation of an OM application prototype for retrieving the general opinions about a phone and its features. Due to the complexity of OM, there is room for the improvement for this task. First, in order to improve the target detection mechanism our intention is to use learning models based on sequence of words (n-gram, Hidden Markov Models, etc.) to detect the topic of published opinion and thus, making a comparative assessment of different techniques, which will be also employed to detect linguistic phenomena based on the consequentiality mechanisms for expressing denials, irony and sarcasm. We also intend to use temporality resolution techniques for detecting lines of argument in different posts from the same source to find possible incoherence and abstract irony or sarcasm. Last but not least, another future work line includes the extension of *EmotiBlog* annotation (data and languages) in order to have at disposal more data for the ML training and test.

References[10]

1. 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.
2. 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.
3. Balahur A., and Montoyo A. 2008. Applying a Culture Dependent Emotion Triggers Database for Text Valence and Emotion Classification. In Proceedings of the AISB

[8] http://webdocs.cs.ualberta.ca/~lindek/minipar.htm

[9] http://lucene.apache.org

[10] included in the paper and for the related work in general

2008 Symposium on Affective Language in Human and Machine, Aberdeen, Scotland.

4. 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.
5. 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.
6. Balahur, A., Boldrini, E., Montoyo, A., Martínez-Barco, P. 2010b. Opinion Question Answering: Towards a Unified Approach. In Proceedings of the ECAI conference.
7. Balahur, A., Boldrini, E., Montoyo, A., Martínez-Barco 2009b. P. Cross-topic Opinion Mining for Realtime Human-Computer Interaction. ICEIS 2009.
8. 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.
9. 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.
10. 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.

11. 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.
12. Chaovalit P, Zhou L. 2005. Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches. In Proceedings of HICSS-05.
13. 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.
14. 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.
15. Dave K., Lawrence S., Pennock, D. 2003. "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews". In Proceedings of WWW-03.
16. 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.
17. Gamon M., Aue S., Corston-Oliver S., Ringger E. 2005. Mining Customer Opinions from Free Text. Lecture Notes in Computer Science.
18. Hatcher E. and Gospodnetic O. 2004. Lucene in Action. Manning Publications.
19. Hatzivassiloglou V., Wiebe J. 2000. Effects of adjective orientation and gradability on sentence subjectivity. In Proceedings of COLING.
20. Liu 2006. Web Data Mining book. Chapter 11
21. Liu, B. (2007). Web Data Mining. Exploring Hyperlinks, Contents and Usage Data. Springer, first edition.
22. Miller, G.A. 1995. WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41
23. Mullen T., Collier N. 2004. Sentiment Analysis Using Support Vector Machines with Diverse Information Sources. In Proceedings of EMNLP.
24. 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.
25. Ohana, B.,Tierney, B. 2009. Sentiment classification of reviews using SentiWordNet, T&T Conference, Dublin Institute of Technology, Dublin, Ireland, 22nd.-23rd.

26. 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.
27. 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.
28. Qiu, G., Liu, B., Bu, J., Chen, C. 2006. Opinion Word Expansion and Target Extraction through Double Propagation. Association for Computational Linguistics
29. 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.
30. Salton, G. and Buckley, C. 1988. Term-weighting Approaches in Automatic Text Retrieval. Inform. Process. Man. 24, 5, 513–523
31. Sebastiani F. 2002. Machine Learning in Automated Text Categorization. ACM Computing Surveys.
32. Strapparava C. Valitutti A. 2004. WordNet-Affect: an affective extension of WordNet. In Proceedings ofthe 4th International Conference on Language Resources and Evaluation, LREC.
33. Turney P. 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. ACL 2002: 417-424.
34. Yang Y. and Pedersen J.O. 1997. A Comparative Study on Feature Selection in Text Categorization. In Proceedings of ICML-97, 14th International Conference on Machine Learning.