EmotiBlog: a finer-grained and more precise learning of subjectivity expression models

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

Alexandra Balahur
University of Alicante, Department of
Software and Computing Systems
abalahur@dlsi.ua.es

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

Andrés Montoyo
University of Alicante, Department of
Software and Computing Systems
montoyo@dlsi.ua.es

Abstract

The exponential growth of the subjective information in the framework of the Web 2.0 has led to the need to create Natural Language Processing tools able to analyse and process such data for multiple practical applications. They require training on specifically annotated corpora, whose level of detail must be fine enough to capture the phenomena involved. This paper presents *EmotiBlog* – a fine-grained annotation scheme for subjectivity. We show the manner in which it is built and demonstrate the benefits it brings to the systems using it for training, through the experiments we carried out on opinion mining and emotion detection. We employ corpora of different textual genres –a set of annotated reported speech extracted from news articles, the set of news titles annotated with polarity and emotion from the SemEval 2007 (Task 14) and ISEAR, a corpus of real-life self-expressed emotion. We also show how the model built from the EmotiBlog annotations can be enhanced with external resources. The results demonstrate that *EmotiBlog*, through its structure and annotation paradigm, offers high quality training data for systems dealing both with opinion mining, as well as emotion detection.

1 Credits

This paper has been supported by Ministerio de Ciencia e Innovación- Spanish Government (grant no. TIN2009-13391-C04-01), and Conselleria d'Educación-Generalitat Valenciana (grant no. PROMETEO/2009/119 and A-COMP/2010/288).

2 Introduction

The exponential growth of the subjective information with Web 2.0 created the need to develop new Natural Language Processing (NLP) tools to automatically process and manage the content available on the Internet. Apart from the traditional textual genres, at present we have new ones such as blogs, forums and reviews. The main difference between them is that the latter are predominantly subjective, containing personal judgments. At the moment, NLP tools and methods for analyzing objective information have a better performance than the new ones the research community is creating for managing the subjective content. The survey called "*The State of the Blogosphere 2009*", published by Technorati[1], demonstrates that users are blogging more than ever. Furthermore, in contrast to the general idea about bloggers, each day it is more and more the number of professionals who decide to use this means of communication, contradicting the common belief about the predominance of an informal editing (Balahur et al., 2009). Due to the growing interest in this text type, the subjective data of the Web is increasing on a daily basis, becoming a reflection of people's opinion about a wide range of topics. (Cui, Mittal and Datar, 2006). Blogs represent an important source of real-time, unbiased information, useful for the development of many applications for concrete purposes. Given the proved importance of automatically processing this data, a new task has appeared in NLP task, dealing with the treatment of subjective data: Sentiment Analysis (SA). The main objective of this paper is to present *EmotiBlog* (Boldrini et al., 2009), a fine-grained annotation scheme for labeling subjectivity in the new textual genres. Subjectivity

[1] http://technorati.com/

Proceedings of the Fourth Linguistic Annotation Workshop, ACL 2010, pages 1–10,
Uppsala, Sweden, 15-16 July 2010. ©2010 Association for Computational Linguistics

can be reflected in text through expressions of emotions beliefs, views (a way of considering something)[2] and opinions, generally denominated "private states" (Uspensky, 1973), not open to verification (Wiebe, 1994). We performed a series of experiments focused on demonstrating that *EmotiBlog* represents a step forward to previous research in this field; its use allows a finer-grained and more precise learning of subjectivity expression models. Starting form (Wiebe, Wilson and Cardie, 2005) we created an annotation schema able to capture a wide range and key elements, which give subjectivity, moving a step forward the mere polarity recognition. In particular, the experiments concern expressions of emotion, as a finer-grained analysis of affect in text and a subsequent task to opinion mining (OM) and classification. To that aim, we employ corpora of different textual genres– a set of annotated reported speech extracted from news articles (denominated JRC quotes) (Balahur et al., 2010) and the set of news titles annotated with polarity and emotion from the SemEval 2007 Task No. 14 (Strapparava and Mihalcea, 2007), as well as a corpus of real-life self-expressed emotion entitled ISEAR (Scherer and Walbott, 1999). We subsequently show, through the quality of the results obtained, that *EmotiBlog*, through its structure and annotation paradigm, offers high quality training for systems dealing both with opinion mining, as well as emotion detection.

3 Motivation and Contribution

The main motivation of this research is the demonstrated necessity to work towards the harmonization and interoperability of the increasingly large number of tools and frameworks that support the creation, instantiation, manipulation, querying, and exploitation of annotated resource. This necessity is stressed by the new tools and resources, which have been recently created for processing the subjectivity in the new-textual genres born with the Web 2.0. Such predominantly subjective data is increasing at an exponential rate (about 75000 new blogs are reported to be created every day) and contains opinions on the most diverse set of topics. Given its worldwide availability, the subjective data on the Web has become a primary source of information (Balahur et al., 2009). As a consequence, new mechanisms have to be implemented so that this data is effectively analyzed and processed. The main challenge of the opinionated content is that, unlike the objective one, which presents facts, the subjective information is most of the times difficult and complex to extract and classify using in grammatically static and fixed rules. Expression of subjectivity is more spontaneous and even if the majority is quite formal, new means of expressivity can be encountered, such as the use of colloquialisms, sayings, collocations or anomalies in the use of punctuation; this is motivated by the fact that subjectivity expression is part of our daily life. For example, at the time of taking a decision, people search for information and opinions expressed on the Web on their matter of interest and base their final decision on the information found. At the same time, when using a product, people often write reviews on it, so that others can have a better idea of the performance of that product before purchasing it. Therefore, on the one hand, the growing volume of opinion information available on the Web allows for better and more informed decisions of the users. On the other hand, the amount of data to be analyzed requires the development of specialized NLP systems that automatically extract, classify and summarize the data available on the Web on different topics. (Esuli and Sebastiani, 2006) define OM as a recent discipline at the crossroads of Information Retrieval and Computational Linguistics, which is concerned not with the topic a document is about, but with the opinion it expresses. Research in this field has proven the task to be very difficult, due to the high semantic variability of affective language. Different authors have addressed the problem of extracting and classifying opinion from different perspectives and at different levels, depending on a series of factors which can be level of interest (overall/specific), querying formula *("Nokia E65"/"Why do people buy Nokia E65?"),* type of text (review on forum/blog/dialogue/press article), and manner of expression of opinion - directly (using opinion statements, e.g. *"I think this product is wonderful!"/"This is a bright initiative"*), indirectly (using affect vocabulary, e.g. *"I love the pictures this camera takes!"/"Personally, I am shocked one can propose such a law!"*) or implicitly (using adjectives and evaluative expressions, e.g. "It's light *as a feather and fits right into my pocket!"*). While determining the overall opinion on a movie is sufficient for taking the decision to watch it or not, when buying a product, people are interested in the individual opinions on the different prod-

[2] http://dictionary.cambridge.org/

uct characteristics. When discussing a person, one can judge and give opinion on the person's actions. Moreover, the approaches taken can vary depending on the manner in which a user asks for the data (general formula such as *"opinions on X" or* a specific question *"Why do people like X?"* and the text source that needs to be queried). Retrieving opinion information in newspaper articles or blogs posts is more complex, because it involves the detection of different discussion topics, the subjective phrases present and subsequently their classification according to polarity. Especially in the blog area, determining points of view expressed in dialogues together with the mixture of quotes and pastes from newspapers on a topic can, additionally, involve determining the persons present and whether or not the opinion expressed is on the required topic or on a point previously made by another speaker. This difficult NLP problem requires the use of specialized data for system training and tuning, gathered, annotated and tested within the different text spheres. At the present moment, these specialized resources are scarce and when they exist, they are rather simplistically annotated or highly domain-dependent. Moreover, most of these resources created are for the English. The contribution we describe in this paper intends to propose solutions to the above-mentioned problems, and consists of the following points: first of all, we overcome the problem of corpora scarcity in other languages except English and also improve the English ones; we present the manner in which we compiled a multilingual corpus of blog posts on different topics of interest in three languages-Spanish, Italian and English. The second issue we tried to solve was the coarse-grained annotation schemas employed in other annotation schema. Thus, we describe the new annotation model, EmotiBlog built up in order to capture the different subjectivity/objectivity, emotion/opinion/attitude aspects we are interested in at a finer-grained level. We justify the need for a more detailed annotation model, the sources and the reasons taken into consideration when constructing the corpus and its annotation. Thirdly, we address an aspect strongly related to blogs annotation: due the presence of "copy and pastes" from news articles or other blogs, the frequent quotes, we include the annotation of both the directly indicated source, as well as the anaphoric references at cross-document level. We discuss on the problems encountered at different stages and comment upon some of the conclusions we have reached while performing this research. this research. Finally, we conclude on our approach and propose the lines for future work.

4 Related Work

In recent years, different researchers have addressed the needs and possible methodologies from the linguistic, theoretical and practical points of view. Thus, the first step involved resided in building lexical resources of affect, such as WordNet Affect (Strapparava and Valitutti, 2004), SentiWordNet (Esuli and Sebastiani, 2006), Micro-WNOP (Cerini et. Al, 2007) or "emotion triggers" (Balahur and Montoyo, 2009). All these lexicons contain single words, whose polarity and emotions are not necessarily the ones annotated within the resource in a larger context. We also employed the ISEAR corpus, consisting of phrases where people describe a situation when they felt a certain emotion. Our work, therefore, concentrates on annotating larger text spans, in order to consider the undeniable influence of the context. The starting point of research in emotion is represented by (Balahur and Montoyo, 2008), who centered the idea of subjectivity around that of private states, and set the benchmark for subjectivity analysis as the recognition of opinion-oriented language in order to distinguish it from objective language and giving a method to annotate a corpus depending on these two aspects – MPQA (Wiebe, Wilson and Cardie, 2005). Furthermore, authors show that this initial discrimination is crucial for the sentiment task, as part of Opinion Information Retrieval (last three editions of the TREC Blog tracks[3] competitions, the TAC 2008 competition[4]), Information Extraction (Riloff and Wiebe, 2003) and Question Answering (Stoyanov et al., 2004) systems. Once this discrimination is done, or in the case of texts containing only or mostly subjective language (such as e-reviews), opinion mining becomes a polarity classification task. Our work takes into consideration this initial discrimination, but we also add a deeper level of emotion annotation. Since expressions of emotion are also highly related to opinions, related work also includes customer review classification at a document level, sentiment classification using unsupervised methods (Turney, 2002), Machine Learning techniques (Pang and Lee, 2002), scoring of features (Dave, Lawrence and Pennock, 2003), using PMI, syntactic relations

[3] http://trec.nist.gov/data/blog.html

[4] http://www.nist.gov/tac/

and other attributes with SVM (Mullena and Collier, 2004), sentiment classification considering rating scales (Pang and Lee, 2002), supervised and unsupervised methods (Chaovalit and Zhou, 2005) and semisupervised learning (Goldberg and Zhou, 2006). Research in classification at a document level included sentiment classification of reviews (Ng, Dasgupta and Arifin, 2006), sentiment classification on customer feedback data (Gamon, Aue, Corston-Oliver, Ringger, 2005), comparative experiments (Cui, Mittal and Datar, 2006). Other research has been conducted in analysing sentiment at a sentence level using bootstrapping techniques (Riloff, Wiebe, 2003), considering gradable adjectives (Hatzivassiloglou, Wiebe, 2000), semisupervised learning with the initial training some strong patterns and then applying NB or self-training (Wiebe and Riloff, 2005) finding strength of opinions (Wolson, Wiebe, Hwa, 2004) sum up orientations of opinion words in a sentence (or within some word window) (Kim and Hovy, 2004), (Wilson and Wiebe, 2004), determining the semantic orientation of words and phrases (Turney and Littman, 2003), identifying opinion holders (Stoyanov and Cardie, 2006), comparative sentence and relation extraction and feature-based opinion mining and summarization (Turney, 2002). Finally, fine-grained, feature-based opinion summarization is defined in (Hu and Liu, 2004) and researched in (Turney, 2002) or (Pang and Lee, 2002). All these approaches concentrate on finding and classifying the polarity of opinion words, which are mostly adjectives, without taking into account modifiers or the context in general. Our work, on the other hand, represents the first step towards achieving a contextual comprehension of the linguistic roots of emotion expression.

5 Corpora

It is well known that nowadays blogs are the second way of communication most used after the e-mail. They are extremely useful and a poll for discussing about any topic with the world. For this reason, the first corpus object of our study is a collection of blog posts extracted from the Web. The texts we selected have distinctive features, extremely different from traditional textual ones. In fact people writing a post can use an informal language colloquialism, emoticons, etc. to express their feelings and it is not rare to find a mix of sources in the same post; people usually mention some facts or discourses and then they give their opinion about them. As we can deduce, the source detection represents one of the most complex tasks. As we mentioned above, we carried out a multilingual research, collecting texts in three languages: Spanish, Italian, and English about three subjects of interest. The first one contains blog posts commenting upon the signing of the Kyoto Protocol against global warming, the second collection consists of blog entries about the Mugabe government in Zimbabwe, and finally we selected a series of blog posts discussing the issues related to the 2008 USA presidential elections. For each of the abovementioned topics, we have gathered 100 texts, summing up a total of 30.000 words approximately for each language. However in this research we start with English but consider as future work labeling the other languages we have. The second corpus we employed for this research is a collection of 1592 quotes extracted from the news in April 2008. As a consequence they are about many different topics and in English (Balahur and Steinberg, 2009). Both of these corpora have been annotated with *EmotiBlog* that is presented in the next section.

6 EmotiBlog Annotation Model

Our annotation schema can be defined as a fine-grained model for labelling subjectivity of the new-textual genres born with the Web 2.0. As mentioned above, it represents a step forward to previous research and it is focused on detecting the linguistic elements, which give subjectivity to the text. The *EmotiBlog* annotation is divided into different levels (Figure 1).

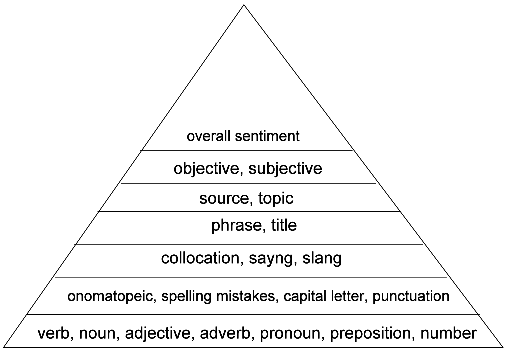
overall sentiment
objective, subjective
source, topic
phrase, title
collocation, sayng, slang
onomatopeic, spelling mistakes, capital letter, punctuation
verb, noun, adjective, adverb, pronoun, preposition, number

Figure 1: General structure of *EmotiBlog*.

As we can observe in Figure 1, the first distinction to be made is between objective and subjective speech. If we are labelling an objective sentence, we insert the source element, while if we are annotating a subjective discourse, a list of elements with the corresponding attributes have to be added. We select among the list of subjective elements and specify the element's attributes. Table 1 presents the annotation model in detail.

Elem.	Description
Obj. speech	Confidence, comment, source, target.
Subj. speech	Confidence, comment, level, emotion, phenomenon, polarity, source and target.
Adjectives	Confidence, comment, level, emotion, phenomenon, modifier/not, polarity, source and target.
Adverbs	Confidence, comment, level, emotion, phenomenon, modifier/not, polarity, source and target.
Verbs	Confidence, comment, level, emotion, phenomenon, polarity, mode, source and target.
Anaphora	Confidence, comment, type, source and target.
Capital letter	Confidence, comment, level, emotion, phenomenon, modifier/not, polarity, source and target.
Punctuation	Confidence, comment, level, emotion, phenomenon, modifier/not, polarity, source and target.
Names	Confidence, comment, level, emotion, phenomenon, modifier/not, polarity, and source.
Phenomenon	Confidence, comment, type: collocation, saying, slang, title, and rhetoric.
Reader Interpretation	Confidence, comment, level, emotion, phenomenon, polarity, source and target.
Author Interpretation	Confidence, comment, level, emotion, phenomenon, polarity, source and target.
Emotions	Confidence, comment, accept, anger, anticipation, anxiety, appreciation, bad, bewilderment, comfort, …

Table 1: *EmotiBlog* structure

Each element of the discourse has its own attributes with a series of features, which have to be annotated. Due to space reasons it is impossible to detail each one of them, however we would like to underline the most innovative and relevant. For each element we are labelling the annotator has to insert his level of confidence. In this way we will assign each label a weight that will be computed for future evaluations. Moreover, the annotator has to insert the polarity, which can be positive or negative, the level (high, medium, low) and also the sentiment this element is expressing. Table 2 presents a complete list of the emotions we selected to be part of *EmotiBlog*. We grouped all sentiments into subgroups in order to help the evaluation process. In fact emotions of the same subgroup will have less impact when calculating the inter-annotation agreement. In order to make this subdivision proper and effective division, we were inspired by (Scherer, 2005) who created an alternative dimensional structure of the semantic space for emotions. The graph below represents the mapping of the term Russell (1983) uses for his claim of an emotion circumflex in two-dimensional valence by activity/arousal space (upper-case terms). As we can appreciate, the circle is divided by 4 axes. Moreover, Scherer distinguishes between positive and negative sentiments and after that between active and passive. Furthermore emotions are grouped between obstructive and conductive, and finally between high power and low power control. We started form this classification, grouping sentiments into positive and negative, but we divided them as high/low power control, obstructive/conductive and active/passive. Further on, we distributed the sentiments within our list into the Scherer slots creating other smaller categories included in the abovementioned general ones. The result of this division is shown in Table 2:

Group	Emotions
Criticism	Sarcasm, irony, incorrect, criticism, objection, opposition, scepticism.
Happiness	Joy, joke.
Support	Accept, correct, good, hope, support, trust, rapture, respect, patience, appreciation, excuse.
Importance	Important, interesting, will, justice, longing, anticipation, revenge.
Gratitude	Thank.
Guilt	Guilt, vexation.
Fear	Fear, fright, troubledness, anxiety.
Surprise	Surprise, bewilderment, disappointment, consternation.
Anger	Rage, hatred, enmity, wrath, force, anger, revendication.
Envy	Envy, rivalry, jealousy.
Indifference	Unimportant, yield, sluggishness.
Pity	Compassion, shame, grief.
Pain	Sadness, lament, remorse, mourning, depression, despondency.
Shyness	Timidity.
Bad	Bad, malice, disgust, greed.

Table 2: Alternative dimensional structures of the semantic space for emotions

Following with the description of the model, we said that the first distinction to be made is between objective and subjective speech. Analysing the texts we collected, we realised that even if the writer uses an objective speech, sometimes it is just apparently objective and for this reason we added two elements: reader and author interpretation. The first one is the impression/feeling/reaction the reader has reading the intervention and what s/he can deduce from the piece of text and the author interpretation is what we can understand from the author (politic orientation, preferences). All this information can be deduced form some linguistic elements that apparently are not so objective as they may appear.

Another innovative element we inserted in the model is the coreference but just at a cross-post level. It is necessary because blogs are composed by posts linked between them and thus cross-document coreference can help the reader to follow the conversations. We also label the unusual usage of capital letters and repeated punctuation. In fact, it is very common in blogs to find words written in capital letter or with no conventional usage of punctuation; these features usually mean shouts or a particular mood of the writer. Using EmotiBlog, we annotate the single elements, but we also mark sayings or collocations, representative of each language. A saying is a well-known and wise statement, which often has a meaning, different from the simple meanings of the words it contains[5]; while a collocation is a word or phrase, which is frequently used with another word or phrase, in a way that sounds correct to native speakers, but might not be expected from the individual words' meanings6. Finally we insert for each element the source and topic. An example of annotation can be: <phenomenon target="Kyoto Protocol" category="phrase" degree="medium" source="w" polarity="positive" emotion="good">The Onion has a <adjective target="Kyoto Protocol" phenomenon="phrase" degree="medium" polarity="positive" emotion="good" source="w" ismodifier="yes">great</adjective> story today titled “Bush Told to Sign Birthday Treaty for Someone Named Kyoto." </phenomenon>

7 Experiments and Evaluation

In order to evaluate the appropriateness of the *EmotiBlog* annotation scheme and to prove that the fine-grained level it aims at has a positive impact on the performance of the systems employing it as training, we performed several experiments. Given that a) *EmotiBlog* contains annotations for individual words, as well as for multi-word expressions and at a sentence level, and b) they are labeled with polarity, but also emotion, our experiments show how the annotated elements can be used as training for the opinion mining and polarity classification task, as well as for emotion detection. Moreover, taking into consideration the fact that *EmotiBlog* labels the intensity level of the annotated elements, we performed a brief experiment on determining the sentiment intensity, measured on a three-level scale: low, medium and high. In order to perform these three different evaluations, we chose three different corpora. The first one is a collection of quotes (reported speech) from newspaper articles presented in (Balahur et al., 2010), enriched with the manual fine-grained annotation of *EmotiBlog*[7]; the second one is the collection of newspaper titles in the test set of the SemEval 2007 task number 14 – Affective Text. Finally, the third one is a corpus of self-reported emotional response – ISEAR (Scherer and Walbott, 1999). The intensity classification task is evaluated only on the second corpus, given that it is the only one in which scores between -100 and 0 and 0 and 100, respectively, are given for the polarity of the titles.

6.1 Creation of training models

For the OM and polarity classification task, we first extracted the Named Entities contained in the annotations using Lingpipe and united through a "_" all the tokens pertaining to the NE. All the annotations of punctuation signs that had a specific meaning together were also united under a single punctuation sign. Subsequently, we processed the annotated data, using Minipar. We compute, for each word in a sentence, a series of features (some of these features are used in (Choi et al., 2005):

- the part of speech (POS)
- capitalization (if all letters are in capitals, if only the first letter is in capitals, and if it is a NE or not)
- opinionatedness/intensity/emotion - if the word is annotated as opinion word, its polarity, i.e. 1 and -1 if the word is positive or negative, respectively and 0 if it is not an opinion word, its intensity (1.2 or 3) and 0 if it is not a subjective word, its emotion (if it has, none otherwise)
- syntactic relatedness with other opinion word – if it is directly dependent of an opinion word or modifier (0 or 1), plus the polarity/intensity and emotion of this word (0 for all the components otherwise)
- role in 2-word, 3-word and 4-word annotations: opinionatedness, intensity and emotion of the other words contained in the annotation, direct dependency relations with them if they exist and 0 otherwise.

We compute the length of the longest sentence in *EmotiBlog*. The feature vector for each of the sentences contains the feature vectors of each of its words and 0s for the corresponding feature vectors of the words, which the current sentence has less than the longest annotated sentence. Finally, we add for each sentence as feature binary features for subjectivity and polarity, the value corresponding to the intensity of opinion and the

[5] Definition according to the Cambridge Advanced Learner's Dictionary

[6] Definition according to the Cambridge Advanced Learner's Dictionary

[7] Freely available on request to the authors.

general emotion. These feature vectors are fed into the Weka[8] SVM SMO ML algorithm and a model is created (EmotiBlog I). A second model (EmotiBlog II) is created by adding to the collection of single opinion and emotion words annotated in EmotiBlog, the Opinion Finder lexicon and the opinion words found in MicroWordNet, the General Inquirer resource and WordNet Affect.

6.2 Evaluation of models on test sets

In order to evaluate the performance of the models extracted from the features of the annotations in *EmotiBlog*, we performed different tests. The first one regarded the evaluation of the polarity and intensity classification task using the *Emoitblog* I and II constructed models on two test sets – the JRC quotes collection and the SemEval 2007 Task Number 14 test set. Since the quotes often contain more than a sentence, we consider the polarity and intensity of the entire quote as the most frequent result in each class, corresponding to its constituent sentences. Also, given the fact that the SemEval Affective Text headlines were given intensity values between -100 and 100, we mapped the values contained in the Gold Standard of the task into three categories: [-100, -67] is high (value 3 in intensity) and negative (value -1 in polarity), [-66, 34] medium negative and [33, 1] is low negative. The values between [1 and 100] are mapped in the same manner to the positive category. 0 was considered objective, so containing the value 0 for intensity. The results are presented in Table 3 (the values I and II correspond to the models EmotiBlog I and EmotiBlog II):

Test Corpus	**Evaluation type**	**Precision**	**Recall**
JRC quotes I	Polarity	32.13	54.09
	Intensity	36.00	53.2
JRC quotes II	Polarity	36.4	51.00
	Intensity	38.7	57.81
SemEval I	Polarity	38.57	51.3
	Intensity	37.39	50.9
SemEval II	Polarity	35.8	58.68
	Intensity	32.3	50.4

Table 3. Results for polarity and intensity classification using the models built from the EmotiBlog annotations

The results shown in Table 2 show a significantly high improvement over the results obtained in the SemEval task in 2007. This is explainable, on the one hand, by the fact that systems performing the opinion task did not have at their disposal the lexical resources for opinion employed in the *EmotiBlog* II model, but also because of the fact that they did not use machine learning on a corpus comparable to *EmotiBlog* (as seen from the results obtained when using solely the *EmotiBlog* I corpus). Compared to the NTCIR 8 Multilingual Analysis Task this year, we obtained significant improvements in precision, with a recall that is comparable to most of the participating systems. In the second experiment, we tested the performance of emotion classification using the two models built using EmotiBlog on the three corpora – JRC quotes, SemEval 2007 Task No.14 test set and the ISEAR corpus. The JRC quotes are labeled using EmotiBlog; however, the other two are labeled with a small set of emotions – 6 in the case of the SemEval data (joy, surprise, anger, fear, sadness, disgust) and 7 in ISEAR (joy, sadness, anger, fear, guilt, shame, disgust). Moreover, the SemEval data contains more than one emotion per title in the Gold Standard, therefore we consider as correct any of the classifications containing one of them. In order to unify the results and obtain comparable evaluations, we assessed the performance of the system using the alternative dimensional structures defined in Table 1. The ones not overlapping with the category of any of the 8 different emotions in SemEval and ISEAR are considered as "Other" and are not included either in the training, nor test set. The results of the evaluation are presented in Table 4. Again, the values I and II correspond to the models EmotiBlog I and II. The "Emotions" category contains the following emotions: joy, sadness, anger, fear, guilt, shame, disgust, surprise.

Test corpus	**Evaluation type**	**Precision**	**Recall**
JRC quotes I	Emotions	24.7	15.08
JRC quotes II	Emotions	33.65	18.98
SemEval I	Emotions	29.03	18.89
SemEval II	Emotions	32.98	18.45
ISEAR I	Emotions	22.31	15.01
ISEAR II	Emotions	25.62	17.83

Table 4. Results for emotion classification using the models built from the EmotiBlog annotations.

The best results for emotion detection were obtained for the "anger" category, where the precision was around 35 percent, for a recall of 19 percent. The worst results obtained were for the ISEAR category of "shame", where precision was around 12 percent, with a recall of 15 per-

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

cent. We believe this is due to the fact that the latter emotion is a combination of more complex affective states and it can be easily misclassified to other categories of emotion. Moreover, from the analysis performed on the errors, we realized that many of the affective phenomena presented were more explicit in the case of texts expressing strong emotions such as “joy” and “anger”, and were mostly related to common-sense interpretation of the facts presented in the weaker ones. As it can be seen in Table 3, results for the texts pertaining to the news category obtain better results, most of all news titles. This is due to the fact that such texts, although they contain a few words, have a more direct and stronger emotional charge than direct speech (which may be biased by the need to be diplomatic, find the best suited words etc.). Finally, the error analysis showed that emotion that is directly reported by the persons experiencing is more “hidden”, in the use of words carrying special signification or related to general human experience. This fact makes emotion detection in such texts a harder task. Nevertheless, the results in all corpora are comparable, showing that the approach is robust enough to handle different text types. All in all, the results obtained using the fine and coarse-grained annotations in *EmotiBlog* increased the performance of emotion detection as compared to the systems in the SemEval competition.

6.3 Discussion on the overall results

From the results obtained, we can see that this approach combining the features extracted from the EmotiBlog fine and coarse-grained annotations helps to balance between the results obtained for precision and recall. The impact of using additional resources that contain opinion words is that of increasing the recall of the system, at the cost of a slight drop in precision, which proves that the approach is robust enough so that additional knowledge sources can be added. Although the corpus is small, the results obtained show that the phenomena it captures is relevant in the OM task, not only for the blog sphere, but also for other types of text (newspaper articles, self-reported affect).

8 Conclusions and future work

Due to the exponential increase of the subjective information result of the high-level usage of the Internet and the Web 2.0, NLP able to process this data are required. In this paper we presented the procedure by which we compiled a multilingual corpus of blog posts on different topics of interest in three languages: Spanish, Italian and English. Further on, we explained the need to create a finer-grained annotation schema that can be used to improve the performance of subjectivity mining systems. Thus, we presented the new annotation model, *EmotiBlog* and justified the benefits of this detailed annotation schema, presenting the sources and the reasons taken into consideration when building up the corpus and its labeling. Furthermore, we addressed the presence of “copy and pastes” from news articles or other blogs, the frequent quotes. For solving this possible ambiguity we included the annotation of both the directly indicated source, as well as the anaphoric references at cross-document level. We performed several experiments on three different corpora, aimed at finding and classifying both the opinion, as well as the expressions of emotion they contained; we showed that the fine and coarse-grained levels of annotation that EmotiBlog contains offers important information on the structure of affective texts, leading to an improvement of the performance of systems trained on it. Although the EmotiBlog corpus is small, the results obtained are promising and show that the phenomena it captures are relevant in the OM task, not only for the blog sphere, but also for other textual-genres. It is well known that OM is an extremely challenging task and a young discipline, thus there is room for improvement above all to solve linguistic phenomena such as the correference resolution at a cross document level, temporal expression recognition. In addition to this, more experiments would need to be done in order to verify the complete robustness of *EmotiBlog*. Last but not least, our idea is to include the existing tools for a more effective semi-supervised annotation. After the training of the ML system we obtain automatically some markables which have to be validated or not by the annotator and the ideal option would be to connect these terms the system detects automatically with tools, such as the mapping with an opinion lexicon based on WordNet (SentiWordNet, WordNet Affect, MicroWordNet), in order to automatically annotate all the synonyms and antonyms with the same or the opposite polarity respectively and assigning them some other elements contemplated into the *EmotiBlog* annotation schema. This would mean an important step forward for saving time during the annotation process and it will also assure a high quality annotation due to the human supervision.

References

Balahur A., Steinberger R., Kabadjov M., Zavarella V., van der Goot E., Halkia M., Pouliquen B., and Belyaeva J. 2010. *Sentiment Analysis in the News.* In Proceedings of LREC 2010.

Balahur A., Boldrini E., Montoyo A., Martínez-Barco P. 2009. *A Comparative Study of Open Domain and Opinion Question Answering Systems for Factual and Opinionated Queries*. In Proceedings of the Recent Advances in Natural Language Processing.

Balahur A., 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., Steinberger R., *Rethinking Sentiment Analysis in the News: from Theory to Practice and back.* In Proceeding of WOMSA 2009. Seville.

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

Boldrini E., Balahur A., Martínez-Barco P., Montoyo A. 2009. *EmotiBlog: a fine-grained model for emotion detection in non-traditional textual genres*. In Proceedings of WOMSA. Seville, Spain.

Boldrini E., Fernández J., Gómez J.M., Martínez-Barco P. 2009. *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.

Carletta J. 1996. *Assessing agreement on classification task: the kappa statistic*. Computational Linguistics, 22(2): 249–254.

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.

Choi Y., Cardie C., Rilloff E., Padwardhan S. 2005. *Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns.* In Proceedings of the HLT/EMNLP.

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

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.

Goldberg A.B., Zhu J. 2006. *Seeing stars when there aren't many stars: Graph-based semi-supervised learning for sentiment categorization.* In HLT-NAACL 2006 Workshop on Textgraphs: Graph-based Algorithms for Natural Language Processing.

Hu M., Liu B. 2004. *Mining Opinion Features in Customer Reviews.* In Proceedings of Nineteenth National Conference on Artificial Intelligence AAAI.

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

Kim S.M., Hovy E. 2004. *Determining the Sentiment of Opinions*. In Proceedings of COLING.

Mullen T., Collier N. 2006. *Sentiment Analysis Using Support Vector Machines with Diverse Information Sources*. In Proceedings of EMNLP. 2004. Lin, W.H., Wilson, T., Wiebe, J., Hauptman, A. "Which Side are You On? Identifying Perspectives at the Document and Sentence Levels". In Proceedings of the Tenth Conference on Natural Language Learning CoNLL.2006.

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.

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

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

Russell J.A. 1983. *Pancultural aspects of the human conceptual organization of emotions*. Journal of Personality and Social Psychology 45: 1281–8.

Scherer K. R. 2005. *What are emotions? And how can they be measured?* Social Science Information, 44(4), 693–727.

Stoyanov V. and Cardie C. 2006. *Toward Opinion Summarization: Linking the Sources*. COLING-ACL. Workshop on Sentiment and Subjectivity in Text.

Stoyanov V., Cardie C., Litman D., and Wiebe J. 2004. *Evaluating an Opinion Annotation Scheme Using a New Multi-Perspective Question and An-*

swer Corpus. AAAI Spring Symposium on Exploring Attitude and Affect in Text.

Strapparava and Mihalcea, 2007 - SemEval 2007 Task 14: Affective Text. In Proceedings of the ACL.

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

Turney P., Littman M. 2003. *Measuring praise and criticism: Inference of semantic orientation from association*. ACM Transactions on Information Systems 21.

Uspensky B. 1973. *A Poetics of Composition*. University of California Press, Berkeley, California.

Wiebe J. M. 1994. *Tracking point of view in narrative*. Computational Linguistics, vol. 20, pp. 233–287.

Wiebe J., Wilson T. and Cardie C. 2005. *Annotating expressions of opinions and emotions in language*. Language Resources and Evaluation.

Wilson T., Wiebe J., Hwa R. 2004. *Just how mad are you? Finding strong and weak opinion clauses*. In: Proceedings of AAAI.

Wiebe J., Wilson T. and Cardie C. 2005. *"Annotation Expressions of Opinions and Emotions in Language*. Language Resources and Evaluation.

Wiebe J., Riloff E. 2005. *Creating Subjective and Objective Sentence Classifiers from Unannotated Texts*. In Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing).