Evaluating EmotiBlog Robustness for SentimentAnalysis Tasks

Javi Fernández, Ester Boldrini, José Manuel Gómez, Patricio Martínez-Barco
University of Alicante –Spain-, GPLSI, Department of Language and Computying Systems.[1]
{Javifm, eboldrini, jmgomez, patricio}@dlsi.ua.es

Abstract. *EmotiBlog* is a corpus labelled with the homonymous annotation schema designed for detecting subjectivity in the new textual genres. Preliminary research demonstrated its relevance as a Machine Learning resource to detect opinionated data. In this paper we compare *EmotiBlog* with the *JRC* corpus in order to check the *EmotiBlog* robustness of annotation. For this research we concentrate on its coarse-grained labels. We carry out a deep ML experimentation also with the inclusion of lexical resources. The results obtained show a similarity with the ones obtained with the *JRC* demonstrating the *EmotiBlog* validity as a resource for the SA task.
Keywords: Sentiment Analysis, EmotiBlog, Machine Learning Experiments.

1 Introduction

The exponential growth of the subjective information on the Web and the employment of new textual genres originated 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). Thus, our research is motivated by the lack of resources, methods and tools to properly treat subjective data. In this paper is we demonstrate that the EmotiBlog annotation schema can be successfully employed to overcome the challenges of SA because of its reliability in terms of annotation. EmotiBlog allows a double level of annotation, coarse and fine-grained and in this paper we test the reliability of the coarse-grained labels. In order to achieve this, we train a Machine Learning (ML) system with two domain-specific corpora annotated with EmotiBlog (EmotiBlog Kyoto[2] and EmotiBlog Phones[3]). We carry out the same experiments with the well-known JRC corpus in order to compare their performances and understand if they are comparable under similar conditions. We carry out a deep study using basic NLP techniques (stemmer, lemmatiser, term selection, etc.) also integrating SentiWordNet (Esuli and Sebastiani, 2006) and WordNet (Miller, 1995) as lexical resources. In previous works

[1] This work has been partially founded by the TEXTMESS 2.0 (TIN2009-13391-C04-01) and Prometeo (PROMETEO/2009/199) projects and also by the complimentary action from the Generalitat valenciana (ACOMP/2011/001).

[2] 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 ones about the Kyoto Protocol (*EmotiBlog Kyoto*). Available on request from authors.

[3] The *EmotiBlog Phones* corpus is composed by users' comments about mobile phones extracted from Amazon UK (http://www.amazon.co.uk). Available on request from authors.

EmotiBlog has been applied successfully to Opinionated Question Answering (OQA) (Balahur et al. 2009 c and 2010a,b), to Automatic Summarization of subjective content (Balahur et al. 2009a), but also to preliminary ML experiments (Boldrini et al. 2010). Thus, the first objective of our research is to demonstrate that our resource is reliable and valuable to train ML systems for NLP applications. As a consequence, our second aim is to show that the next step of our research will consist in a deeper text classification for the SA task –and its applications-. In fact, alter having demonstrated the validity of the coarse-grained annotation, we will test the EmotiBlog fine-grained annotation. We believe there is a need for positive/negative text categories, but also emotion intensity (high/medium/low), emotion type (Boldrini et al, 2009a) and the annotation of the elements that give the subjectivity to the discourse, contemplated by EmotiBlog, not just at sentence level.

2 Related Work

The first step of SA consists 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. The starting point of research in emotion is represented by (Wiebe 2004), who focused the idea of subjectivity around that of private states setting the benchmark for subjectivity analysis. Furthermore, authors show that the discrimination between objective/subjective discourses is crucial for the sentiment task, as part of Opinion Information Retrieval (last three editions of the TREC Blog tracks[4] competitions, the TAC 2008 competition[5]), Information Extraction (Riloff and Wiebe, 2003) and QA (Stoyanov et al., 2005) systems. Related work also includes customer review classification at a document level, 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), on customer feedback data (Gamon, Aue, Corston-Oliver, Ringger, 2005). Other research has been conducted in analysing sentiment at a sentence level using bootstrapping techniques (Riloff, Wiebe, 2003), considering gradable adjectives (Hatzivassiloglou, Wiebe, 2000), (Kim and Hovy, 2004), or determining the semantic orientation of words and phrases (Turney and Littman, 2003). Other work includes (Mcdonald et al. 2007) who investigated a structured model for jointly classifying the sentiment of a text at varying levels of granularity. Neviarouskaya (2010) classified texts using finegrained attitude labels basing its work on the compositionality principle and an approach based on the rules elaborated for semantically distinct verb classes, while 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. They divide the emotion classification into two steps: sentiment polarity and emotion classification. Our work starts from the conclusions drawn by (Boldrini

[4] http://trec.nist.gov/data/blog.html
[5] http://www.nist.gov/tac/

et al 2010) in which authors performed several experiments on three different corpora, aimed at finding and classifying both the opinion, as well as the expressions of emotion they contained. They 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. Thanks to EmotiBlog- annotated at sentence, as well as element level- we have the possibility of carrying out ML experiments of different nature proposing a mixed approach based on the ML training but also the lexical resources integration.

3 Training Corpora

The corpora (in English) we employed for our experiments are EmotiBlog Kyoto extended with the collection of mobile phones reviews extracted from Amazon (EmotiBlog Phones). It allows the annotation at document, sentence and element level (Boldrini et al. 2010), distinguishing between objective and subjective discourses. The list of tags for the subjective elements is presented in (Boldrini et al, 2009a). We also use 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. For all of these elements, the common attributes are annotated: polarity, degree and emotion. As we want to compare the two corpora, we will consider entire sentences and evaluate the polarity, to adapt to the JRC annotation schema. Table 1 presents the size of all the corpora in sentences divided by its classification.

Table 1. Corpora size in sentences.

	EB Kyoto	**EB Phones**	**EB Full**	**JRC**
Objective	347	172	519	863
Subjective	210	246	456	427
Positive	62	198	260	193
Negative	141	47	188	234
Total	557	418	975	1290

4 Machine Learning Experiments and Discussion

In order to demonstrate that *EmotiBlog* is valuable resource for ML, we perform a large number of experiments with different approaches, corpus elements and resources. As features for the ML system we use the classic *bag of words* initially. To reduce the dimensionality we also employ techniques such as *stemming*, *lemmatization* and *dimensionality reduction by term selection* (TSR) methods. For TSR, we compare two approaches, *Information Gain* (IG) and *Chi Square* (X2), because they reduce the dimensionality substantially with no loss of effectiveness (Yang and Pedersen, 1997). For weighting these features we evaluate the most common methods: *binary weighting*, *tf/idf* and *tf/idf normalized* (Salton and Buckley, 1988). As supervised learning method we use *Support Vector Machines* (SVM) due to its good results in text categorization (Sebastiani, 2002) and the promising results obtained in previous studies (Boldrini et al. 2009b). We also evaluate if grouping features by their semantic relations increases the coverage in the test corpus and reduces 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

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

would introduce noise in the evaluation and thus a low performance. But we believe that if we include all senses of a term in the set of features the TSR will choose only the correct ones. For example, using all *WordNet* (WN) senses of each term as learning features, the TSR methods could remove the non-useful senses to classify a sample in the correct class. In this case this disambiguation methods would be adequate. As lexical resources for these experiments we employ WN, but also *SentiWordNet* (SWN), since the use of this specific OM resource demonstrated to improve the results of OM systems. 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 *Spositive* with their negative and positive scores respectively. The experiments with lexical resources have been carried out with five different configurations using: **i)** only SWN synsets (experiment *s*), **ii)** only WN synsets (experiment *w*), **iii)** both SWN and WN synsets (experiment *sw*), **iv)** only SWN synsets including sentiment scores (experiment *ss*) and **v)** both SWN and WN synsets including sentiment scores (experiment *sws*). In case a term is not found in any of the lexical resources, then its lemma is left. Moreover, to solve the ambiguity, two techniques have been adopted: including all its senses and let the TSR methods perform the disambiguation (experiments *s*, *w*, *sw*, *ss* and *sws*), or including only the most frequent sense for each term (experiments *s1*, *w1*, *sw1*, *ss1* and *sws1*). We made an exhaustive combination of all the possible parameters (tokenization, weighting, feature selection and use of lexical resources), with the different classifications and corpus, which is summarized in Table 2, where we only show the best results for each pair classification-corpus because of space reasons.

Table 2. Best configuration and result for each pair classification-corpus.

	EB Kyoto		**EB Phones**		**EB Full**		**JRC**	
	Conf	F1	Conf	F1	Conf	F1	Conf	F1
Objectivity	sws	0.6647	sws1	0.6405	sw	0.6274	w1	0.6088
Polarity	ss1	0.7602	ss	0.**8093**	ss1	0.6374	w1	0.5340

The lower results belong to the *JRC* and *EmotiBlog Full* corpora, although they are the bigger ones. They are not domain-specific so is more difficult for the ML system to create a specialized model. But their best results using similar techniques are ver similar. This fact shows us that the annotation schema and process is valid. On the other hand, experiments with *EmotiBlog Kyoto* and *Phones* obtain the best results because they are domain-specific. This makes *EmotiBlog* more usable in real-word applications, which demand higher performance and usually belong to a specific domain. Regarding the evaluation of the different techniques, we can see that the best results include the lexical resources. Having a look to the totality of experiments, they are always in the top positions. Moreover, in Table 2 we can see that SWN is present in 6 of the 8 best results shown, and the sentiment scores in 5 of them. This encourages us to continue using SWN in our following experiments and find new ways to take advantage of the sentiment information it provides. The other mentioned techniques (tokenization, weighting and feature selection) affect the results but not in significant way. Their improvements are very small and do not seem to follow any pattern. As future work we propose to experiment with other well-known corpora and combinations of different ones, in order to evaluate if the improvements depend on the type and the size of the corpus. We will also continue evaluating the EmotiBlog robustness. Specifically we will test the reliability of its fine-grained annotation.