Machine Learning Techniques for Automatic Opinion Detection in Non-Traditional Textual Genres

Ester Boldrini*, Javi Fernández•, José M. Gómez* and Patricio Martínez-Barco*

Depto. Lenguajes y Sistemas Informáticos
Universidad de Alicante,
Carretera San Vicente del Raspeig s/n - 03690 San Vicente del Raspeig – Alicante,
{eboldrini,jmgomez,patricio}@dlsi.ua.es*
javier.fernandez@eltallerdigital.com•

Abstract This paper presents a preliminary study in which Machine Learning experiments applied to Opinion Mining in blogs have been carried out. We created and annotated a blog corpus in Spanish using EmotiBlog. We evaluated the utility of the features labelled firstly carrying out experiments with combinations of them and secondly using the feature selection techniques, we also deal with several problems, such as the noisy character of the input texts, the small size of the training set, the granularity of the annotation scheme and the language object of our study, Spanish, with less resource than English. We obtained promising results considering that it is a preliminary study.

Keywords: Opinion Mining, Sentiment Analysis, Machine Learning, Blogs, Emotion Annotation-Scheme, Feature Selection.

1 Introduction

Opinion Mining (OM), also called Sentiment Analysis, can be defined as the extraction of a sentiment from an unstructured source such as text, images or audio. The recognized sentiments can be classified as positive or negative or a more fine-grained sentiment classification scheme can be used [1].
With the rapid expansion of the Social Web, users have become comfortable with the Internet generating a wide growth of web-culture communities such as social-networking sites, forums or blogs. By means of these new channels of communication they share their private opinions about whatever is on their mind. The result is a huge amount of precious data with a great potential but already unexploited. It could be

exploited to carry out relevant studies for different entities, depending on the application scenario. One of them could be the economic framework. In this case the data could be employed to carry out studies about our products or competitors' ones, market researches or economic crisis prevention. Another relevant scenario of application could be the social framework: the data could be exploited from delinquency to suicide prevention or for reconstructing crimes.

OM offers several advantages. Firstly, the users, in the blogosphere, usually express their opinions in a more pronounced way than in a face-to-face conversation or different situations and they are also influencing each other reading such posts. Secondly, opinions are extracted in real-time, allowing quicker response times to market changes, for example. Last but not least, Information Retrieval OM helps in discriminating opinionated documents from the objective ones.

This article presents some preliminary experiments we carried out in order to determinate the effectiveness of EmotiBlog [2 and 3], an annotation scheme for emotions detection in non-traditional textual genres.

The article is structured as follows: Section 2 presents related work in this area; Sections 3 and 4 show the motivation of this research together with our preliminary experiments; Section 5 draws a discussion; and, finally Section 6 expresses our conclusions and future work.

2 Related work

Researches as well as commercial companies have increased their interests in OM. [4 and 5] have been the pioneers in researching into text classification according to sentiment or textual genre (a text produced according to a particular model or style) [6]. Moreover, current available systems identify the opinion at a sentence or document level [30]. [7 and 8] analyse other aspects of the content taking into account the sentiment type such as "*happy, sad, anger, fear, disgust, surprise"*. In OM certain sentiments are expressed in two or more words and the accurate detection of negation is important since it reverses the polarity. In an alternative approach to negation [10], each word following a negation until the first punctuation receives a tag indicating negation to the learning algorithm. [11 and 12] select only a subset of the words, often by considering only adjectives detected with a part-of-speech (POS) recognizer. [13] constructs an opinion word lexicon, whereas [14] identifies words describing the features of a product and use them in the classification.

It is worth mentioning that supervised techniques are widely applied for recognizing the sentiment of complete documents [15]. An example of sentence classification is found in [16]. However opinion extraction from noisy Web texts (such as blogs) is still a challenging task. The blog texts present a large number of anomalies: first, on a lexical level, new words, contractions of existing words, and community jargon are no exception; and, second, on a syntactic level, we often find anomalies such as punctuation or sentence structure, which cannot be considered real sentences.

On the one hand, the TREC 2006 Blog track[1] aimed to explore the information seeking behavior in the blogosphere and contained an opinion retrieval task. On the other hand, OM in legal blogs has been performed [17], mood levels also were detected in blog posts [18] and six basic emotion categories were recognized in blog sentences [19]. However these studies did rarely involve the processing of real malformed language (a frequent case in blog posts). The manual annotation of sufficient and representative training examples still remains a problem. In an active learning approach all the examples are labeled by a human, but the limited set of examples to be labeled is carefully selected by the machine. As a consequence of the lack of annotated training examples, it could be very useful to transfer a learned classification model to a new domain or new language. This is interesting to see the effect of using annotated examples of one subject domain (e.g., car opinions) for training a classifier that is used on another domain (e.g., movies). This problem of cross-domain and cross-language issue is already tackled by Finn [22 and 23]. Overall, they show that OM is a very domain-specific problem and it is hard to create a domain independent classifier. It is worth mentioning that this model is extremely fine-grained and due to this reason Machine Learning (ML) [9] experiments are also the best way to tackle the problem of the equilibrium between granularity and effectiveness. In fact, with our feature selection we will be able to improve the model, eliminating all the irrelevant elements that could generate noise obtaining an effective and fine-grained model for OM in non-traditional genres. The automated categorization of texts into predefined categories has witnessed a booming interest in the last 10 years, due to the increased availability of documents in digital form and the ensuing need to organize them. In the research community the dominant approach to this problem is based on ML techniques: a general inductive process automatically builds a classifier by learning, from a set of pre-classified documents, the features of the categories. The advantage of reducing these features is very good effectiveness, considerable savings in terms of expert labor power and straightforward portability to different domains. Examples of researches in this area could be [26 and 27]. They use sophisticated information-theoretic functions such as information gain.

3 Motivation

At the time of taking a decision, an increasing number of people search for information and opinions expressed on the Web on their subject of interest basing their final decision on the information found. On the one hand, the growing volume of opinion information available on the Web allows for better and more informed decisions of the users but, on the other hand, the quantity of data to be analyzed imposed the development of specialized Natural Language Processing (NLP) systems able to automatically extract, classify and summarize the opinions on the Web. Research on OM, has proven that this task is extremely complex, due to the high semantic variability of affective language. In some researches mentioned in Section 2, there were previous approaches for corpus annotation. They mostly concentrated on

[1] http://trec.nist.gov/tracks.html

subjectivity versus objectivity classification more than annotating emotion on a fine-grained scale. The annotation model we employ includes word/phrase/text levels of annotation and, thus, it is useful for constructing similarity models for the training of ML algorithms working with different values of n-grams, as well as sentences as a whole. In fact EmotiBlog is an annotation scheme for emotion detection in non-traditional textual genres. The first contribution of our paper is to use a fine-grained model that allows annotating deeply our collection of blog posts. In fact, EmotiBlog can be used either for basic tasks of sentiment polarity classification, as well as emotion detection, either in very fine-grained categories and, also, psychology-based emotion classes. Another contribution this paper brings is the creation and annotation of a corpus composed by several blogs in Spanish as explained in Section 4. One of our purposes is to concentrate our work on this language, that has less resources than English. Furthermore, most work done in OM only concentrated on classifying polarity of sentiments into positive or negative. Thus, another contribution our work brings is the classification according to three categories: positive, negative and neutral, improving previous work done and also identifying the words that characterize the polarity. The last category is composed by both opinionated sentences in which there is no clear indication of approval or disapproval of an idea and objective sentences. In this way the entire corpus is annotated. Finally, we would like to say that a part of the widely-used ML experiments, we also carry out a feature selection study to understand which elements of the model are the most relevant and produce better.

4 Corpus

The corpus we used for our experiments is composed by texts in Spanish and has been collected from blogs. We decided to use this new textual genre due to its wide diffusion on the Web 2.0. We are convinced that blogs offer a precious opportunity to carry out interesting and useful researches focused on concrete applications. Generally blogs are considered to be written in an informal language, but [29] demonstrated that they are written by a mixture of styles. As a consequence, studying the features of its language in most cases results extremely complex due to this heterogeneity. The blog corpus we collected is composed by posts about the Kyoto Protocol. We decided to select texts about this subject due to the interest this topic raises in users. It is worth mentioning that we selected texts in Spanish in order to start giving importance to languages with less resource than English. Since it is a preliminary research, we do not have at our disposal a balanced set of positive, negative and neutral sentences and the positive polarity is the one with less representatives. The collection is composed by real blog posts and, thus, this difference reflects people's opinion about the topic. After having collected the corpus we labelled it using EmotiBlog, a fine-grained annotation scheme for emotions in non-traditional textual genres [31] for Spanish, Italian and English. **Figure 1** shows its structure.

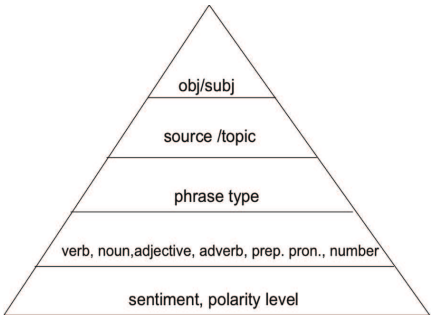

Figure 1: EmotiBlog annotation scheme

This is a visual representation of EmotiBlog. As we can see, the first distinction to be done is between objective or subjective discourse. After that, we insert the source and the topic, the phrase type and the possible elements that express subjectivity. Finally the sentiment type together with its polarity level are annotated. This model is extremely fine-grained and in some cases this could generate noise at the time of making ML experiments; there are many elements and eliminating some of them would mean a significant improvement always taking into account that we are also convinced that we need a fine-grained annotation scheme due to the fact that previous works only focused on the polarity at a document or sentence level. The purpose of our research is to offer a model that enables a more deep study of language used in the new textual genres.

5 Experiments

In order to evaluate the consistency of our model and to discover elements that generate noise we carried out a set of preliminary ML experiments. We used each sentence of the corpus as an individual instance in a classification task. The annotation elements of each of these sentences were treated as features. Finally, we consider each sentence polarity as a category, thus as to evaluate the effectiveness of each of our experiments employing the well-known 10-fold cross validation evaluation method. In the following sections we describe these evaluations in detail. We would like to add that we used the Weka[2] implementation of two widely used ML algorithms: Support Vector Machines (SVM) [24 and 27] and Multinomial Naïve Bayes (MNB) [25 and 27]. We chose the first one due to its robustness against noise and the second because of its simplicity and efficiency. Moreover [27] demonstrated their effectiveness in this kind of tasks. In some experiments we also used the Snowball[3] stemmer for the Spanish language because of its efficiency and availability.

[2] http://www.cs.waikato.ac.nz/ml/weka/
[3] http://snowball.tartarus.org/

5.1 Preliminary experiments

As starting point we extracted a bag of word list from the corpus, and we decided to include in it only the annotated words and phrases as features for the ML system. Moreover, since Spanish is a language with many accents, users in blogs, often forget to put them. For this reason we decided to eliminate them in order to improve the system recall. Furthermore, stopwords are terms with little information [28], and as consequence we opted for deleting all of them from the bag of words in some experiments in order to observe how their inclusion or exclusion can affect the results. While analyzing the corpus we realized that the negation, depending on its situation in the sentence, can reverse the polarity and thus it could be a relevant feature. Therefore we carried out two evaluations using the stopword list: one including the negation in the stopword list and the other without it. During the labelling process we also took into consideration cases of sayings and collocations. They have high subjective connotation and they are extremely representative of a culture. In this initial research we considered them simply as language features that infer subjectivity in texts. However, we will exploit them in order to carry out pragmatic studies. Sayings and collocations are labelled as a whole because their meaning is different from the simple meanings of the words it contains. Once we performed the abovementioned experiments, we also repeated them including a stemmer, to improve the recall.

Table 1 presents the results obtained in our experiments, being WB our starting point. In WB_{SW} we eliminate all the stopwords including the negation, whereas in WB_{SW+NO} we maintain it. WB_{PH} considers all the words that compose a saying or a collocation as a single feature. The next experiments repeat the previous ones but applying stemmer (WB_{ST}, WB_{SW+ST}, $WB_{SW+NO+ST}$, WB_{PH+ST}).

The results show that our starting point reaches an accuracy of 68.5% and a f-measure of 64.4% using SVM. With MNB the results are about 6% and 10% worse in terms of accuracy and f-measure respectively.

The best results were obtained using the Snowball stemmer achieving an accuracy of 71.4% and a f-measure of 68.3%. In fact the use of stemmer improves the results in any experiment we performed with it.

As we can see in **Table 1** the use of a full list of stopwords influences negatively the results. This is probably due to some stopwords can be useful to detect the polarity of a sentence.

Table 1: Results in terms of accuracy and f-measure for the MNB and SVM machine learning algorithms for each option

	# Features	MNB		SVM	
		Accuracy	**F-measure**	**Accuracy**	**F-measure**
WB	941	0,647	0,592	0,685	0,644
WB_{SW}	877	0,532	0,420	0,625	0,572
WB_{SW+NO}	878	0,566	0,477	0,654	0,610
WB_{PH}	875	0,588	0,511	0,663	0,620
WB_{ST}	819	**0,672**	**0,625**	**0,714**	**0,683**
WB_{SW+ST}	764	0,594	0,516	0,661	0,618
$WB_{SW+NO+ST}$	765	0,622	0,556	0,689	0,652
WB_{PH+ST}	781	0,617	0,554	0,694	0,659

Another aspect to underline is that the negation is an important feature due to its inclusion in the features list raises the results between 4% and 7% with respect to the experiment without this feature.
We would like to conclude that the results obtained are promising and they encourage us to continue improving the model. The first reason of this could be the size of the corpus. In fact, we worked with a little collection of texts and as a consequence the number of examples of positive, negative and neutral sentences was extremely different.

5.2 Feature selection

In order to evaluate the features described in the last section, we also made experiments using the Information Gain (IG) [26] feature selection algorithm. The goal of this dimensionality reduction is to obtain the best features for the polarity classification. We employ the global feature selection (instead of the local one) which measures the relevance of each term of the corpus taking into account the three categories of polarity. The results are shown on ***Table 2***.

Table 2: Results in terms of accuracy and f-measure for the MNB and SVM Machine Learning algorithms for each reduction

% Reduction	# Features	MNB		SVM	
		Accuracy	F-Measure	Accuracy	F-Measure
80%	752	0,487	0,322	0,496	0,341
85%	799	0,485	0,320	0,525	0,399
90%	846	0,488	0,325	0,539	0,429
95%	893	0,510	0,371	0,598	0,528
99%	931	0,564	0,470	0,650	0,602

As we can see in ***Table 2*** the results are becoming worst when reducing the number of features with the IG method. Studying these results we observed that the feature selection algorithm removes primarily features representing positive sentiment because of the small quantity of them. From these results we can deduce that applying global feature selection on an unbalanced small corpus does not improve the previous evaluation.

6 Conclusions and future work

In this paper we presented a corpus composed by blog posts in Spanish about the Kyoto protocol. It has been labeled using EmotiBlog, an annotation scheme for emotion detection in non-traditional textual genres. Moreover, we carried out some ML experiments and a feature selection process obtaining low results due to the small size of the corpus and to the extreme granularity of the model. As a consequence we

will employ a binary classification instead of a multi-label one in order to solve these problems. We also contemplate the possibility of using a local feature selection instead of the global one, to extract the best features for each polarity and other well-known feature selection. The results show that using only a stemmed bag of words with a SVM classifier will reach the best performance. Our purpose is to explore alternative ML algorithms with different features. We presented the evaluation together with the results obtained and solutions to improve percentages. There is no doubt about the fact that opinion labeling is a very complex task and above all in the new textual genres that are written in a mixture of styles, also with grammar mistakes. As a future work we will also study the effect of including linguistic tools such as WSD systems or lemmatizers, together with a study focused on finding what are the most relevant stopwords to be included in our experiments, and also to contemplate more complex negation structures. It is worth mentioning that the binary classification will also be effective in determining in which kind of polarity the negation is determinant. Finally, another relevant study we will carry out is the integration of semantic roles in order to detect the source of the discourse and also the Name Entity Recognition, useful to understand which is the most relevant topic of our documents. We will also collect a larger corpus and label it using a new version of EmotiBlog, the result of all the improvements after a careful research.

Acknowledgements

This paper has been supported by the next projects: "Question Answering Learning technologies in a multiLingual and Multimodal Environment (QALL-ME)" (FP6 IST-033860) and "Intelligent, Interactive and Multilingual Text Mining based on Human Language Technologies (TEXT-MESS)"(TIN2006-15265-C06-01).

References

1. Erik Boiy and Marie-Francine Moens. A machine learning approach to sentiment analysis in multilingual Web texts. Inf Retrieval (2009) 12:526–558. 2009
2. Ester Boldrini, Alexandra. Balahur, Patricio Martínez-Barco, Andrés Montoyo. EmotiBlog: an Annotation Scheme for Emotion Detection and Analysis in Non-traditional Textual Genres. In Proceedings of the 5th International Conference on Data Mining. Las Vegas, Nevada, USA. 2009.
3. E. Boldrini, A. Balahur, P. Martínez-Barco, A. Montoyo. EmotiBlog: an Annotation Scheme for Emotion Detection and Analysis in Non-traditional Textual Genres. In Proceedings of the 5th International Conference on Data Mining. Las Vegas, Nevada, USA. 2009.
4. Hearst, M. A. (1992). Direction-based text interpretation as an information access refinement. In P. Jacobs (Ed.), Text-based intelligent systems: Current research and practice in information extraction and retrieval (pp. 257–274). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
5. Kessler, B., Nunberg, G., & Schütze, H. (1997). Automatic detection of text genre. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics (pp. 32–38). Somerset, NJ: Association for Computational Linguistics.

6. Argamon-Engelson S., Koppel, M. and Avneri, G. (1998) Style-based Text Categorization: What Newspaper Am I Reading? In proceedings of AAAI workshop.
7. Huber, R., Batliner, A., Buckow, J., Nöth, E., Warnke, V., & Niemann, H. (2000). Recognition of emotion in a realistic dialogue scenario. In Proceedings of the International Conference on Spoken Language Processing (Vol. 1, pp. 665–668). Beijing, China.
8. Turney, P. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of ACL-02, 40th Annual Meeting of the Association for Computational Linguistics (pp. 417–424). Philadelphia, PA: Association for Computational Linguistics.
9. Polanyi, L., & Zaenen, A. (2006). Contextual valence shifters. In J. Shanahan, Y. Qu, & J. Wiebe (Eds.), Computing attitude and affect in text: Theory and applications (pp. 1–10). Springer.
10. Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of ACL-04, 42nd Meeting of the Association for Computational Linguistics (pp. 271–278). East Stroudsburg, PA: Association for Computational Linguistics.
11. 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 (pp. 79–86). Philadelphia, PA: Association for Computational Linguistics.
12. Wiebe, J. (2000). Learning subjective adjectives from corpora. In Proceedings of AAAI-00, 17th Conference of the American Association for Artificial Intelligence (pp. 735–740). Austin, TX: AAAI Press/The MIT Press.
13. Riloff, E., Wiebe, J., & Wilson, T. (2003). Learning subjective nouns using extraction pattern bootstrapping. In Proceedings of CoNLL-03, 7th Conference on Natural Language Learning (pp. 25–32). Edmonton, CA.
14. Hu, M., & Liu, B. (2004). Mining opinion features in customer reviews. In Proceedings of AAAI-04, 19th National Conference on Artificial Intellgience (pp. 755–760). San Jose, USA.
15. Mullen, T., & Collier, N. (2004). Sentiment analysis using support vector machines with diverse information sources. In Proceedings of EMNLP-04, 9th Conference on Empirical Methods in Natural Language Processing (pp. 412–418). Barcelona, Spain.
16. Dave, K., Lawrence, S., & Pennock, D. M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of WWW-03, 12th International Conference on the World Wide Web (pp. 519–528). New York: ACM Press.
17. Conrad, J. G., & Schilder, F. (2007). Opinion mining in legal blogs. In Proceedings of ICAIL'07, 11th International Conference on Artificial Intelligence and Law (pp. 231–236). New York: ACM.
18. Mishne, G. (2005). Experiments with mood classification in blog posts. In Style2005, 1st Workshop on Stylistic Analysis of Text for Information Access at SIGIR 2005.
19. Aman, S., & Szpakowicz, S. (2008). Using Roget's thesaurus for fine-grained emotion recognition. In Proceedings of the International Joint Conference on NLP (IJCNLP) (pp. 296–302).
20. Iyengar, V. S., Apte, C., & Zhang, T. (2000). Active learning using adaptive resampling. In Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 92–98). New York: ACM.
21. Roy, N., & McCallum, A. (2001). Toward optimal active learning through sampling estimation of error reduction. In Proceedings of ICML-01, 18th International Conference on Machine Learning (pp. 441–448). San Francisco, CA: Morgan Kaufmann.

22. Finn, A., & Kushmerick, N. (2003). Learning to classify documents according to genre. Journal of the American Society for Information Science, 57, 1506–1518. Special issue on Computational Analysis of Style.
23. Gamon, M. (2004). Sentiment classification on customer feedback data: Noisy data, large feature vectors, and the role of linguistic analysis. In Proceedings of COLING-04, the 20th International Conference on Computational Linguistics (pp. 841–847). Geneva, CH.
24. Vapnik, V. (1995). The nature of statistical learning theory. Springer-Verlag New York.
25. Lewis, D., & Gale, W. (1994). A sequential algorithm for training text classifiers. In Proceedings of the 17th Annual international ACM SIGIR Conference on Research and Development in Information Retrieval. Springer-Verlag New York.
26. Yang, Y., & Pedersen, J. (1997). A comparative study on feature selection in text categorization. Proceedings of ICML-97, 14th International Conference on Machine Learning.
27. Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys (CSUR), 34, 1-47.
28. Zipf, G. (1935). The psychobiology of language. Houghton-Mifflin.
29. M.,Tavosanis. Linguistic features of Italian blogs: literary language. New Text. Wikis and blogs and other dynamic text sources, pp 11-15, Trento,vol. 1, 2006.
30. J. Wiebe, T. Wilson, and C. Cardie Annotating expressions of opinions and emotions in language. Language Resources and Evaluation, volume 39, issue 2-3, pp. 165-210, 2005.
31. A. Balahur, E. Boldrini, A. Montoyo, P. Martínez-Barco. Fact versus Opinion Questions Classification and Answering: Challenges and Keys. In ICAI'09 - The 2009 International Conference on Artificial Intelligence. Las Vegas, Nevada, USA. 2009.