SINAI: Machine Learning and Emotion of the Crowd for Sentiment Analysis in Microblogs

E. Martínez-Cámara
SINAI research group
University of Jaén
E-23071, Jaén (Spain)
emcamara@ujaen.es

A. Montejo-Ráez
SINAI research group
University of Jaén
E-23071, Jaén (Spain)
amontejo@ujaen.es

M. T. Martín-Valdivia
SINAI research group
University of Jaén
E-23071, Jaén (Spain)
maite@ujaen.es

L. A. Ureña-López
SINAI research group
University of Jaén
E-23071, Jaén (Spain)
laurena@ujaen.es

Abstract

This paper describes the participation of the SINAI research group in the 2013 edition of the International Workshop SemEval. The SINAI research group has submitted two systems, which cover the two main approaches in the field of sentiment analysis: supervised and unsupervised.

1 Introduction

In the last years, the sentiment analysis (SA) research community wants to go one step further, which consists in studying different texts that usually can be found in commerce websites or opinions websites. Currently, the users publish their opinions through other platforms, being one of the most important the microblogging platform Twitter[1]. Thus, the SA research community is focused on the study of opinions that users publish through Twitter. This interest is shown in several workshops focused on the study of SA in Twitter:

1. RepLab 2012 at CLEF[2] (Amigó et al., 2012): Competition carried out within the CLEF conference, where the participants had to develop a system for measuring the reputation of commercial brands.
2. TASS 2012 at SEPLN[3](Villena-Román et al., 2013): Satellite event of the SEPLN 2012 Conference to foster the research in the field of SA in social media, specifically focused on the Spanish language.

In this paper is described the participation of the SINAI[4] research group in the second task of the 2013 edition of the International Workshop SemEval (Wilson et al., 2013). We have submitted two systems (constrained and unconstrained). The constrained system follows a supervised approach, while the unconstrained system is based on an unsupervised approach which used two linguistic resources: the Sentiment Analysis Lexicon[5] (Hu and Liu, 2004) and WeFeelFine[6] (Kamvar and Harris, 2011).

The paper is organized as follows: first we present a description of the preparing data process. Then the constrained system is outlined. The participation overview finishes with the description of the unconstrained system.

2 Preparing data

The organizers provided two sets of data, one for training and another for the development. The data was concerned by a set of identification number of tweets with their corresponding polarity label. We used the script provided by the organizers to download the two sets of tweets.

[1] http://twitter.com

[2] http://limosine-project.eu/events/replab2012

[3] http://www.daedalus.es/TASS/

[4] http://sinai.ujaen.es

[5] http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar

[6] http://wefeelfine.org

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 402–407, Atlanta, Georgia, June 14-15, 2013. ©2013 Association for Computational Linguistics

The python script was no able to download all the tweets. The training set was composed by 8,633 tweets and the development set by 1,053 tweets.

The data preparation is a step in the workflow of most data mining tasks. Also, in Natural Language Processing is usual the preparation of the documents or the texts for their further processing. Internet is usually the source of texts for SA tasks, so the application of a specific processing to those texts with the aim of extracting their polarity is recommended. The texts published in Twitter have several issues that must be resolved before processing them:

1. The linguistic style of tweets is usually informal, with a intensive usage of abbreviations, idioms, and jargon.
2. The users do not care about the correct use of grammar, which increases the difficulty of carrying out a linguistic analysis.
3. Because the maximum length of a tweet is 140 characters, the users normally refer to the same concept with a large variety of short and irregular forms. This problems is known as data sparsity, and it is a challenge for the sentiment-topic task.
4. The lack of context, which makes difficult to extract the semantics of these sort pieces of text.

Before applying a cleaning process to the corpus with the aim of overcoming the issues described above, we have studied the different kinds of marks, like emoticons, question and exclamation marks or hashtags in the tweets.

Regarding the issues listed above and the marks in the tweets, we have carried out a cleaning and a normalization process which imply the following operations:

1. The uppercase characters have been exchanged by lowercase characters.
2. Links have been replaced by the token “_ULR_”.
3. Question and exclamation marks have been switched to the tokens “_QUESTION_” and “_EXCLAMATION_” respectively.
4. Mentions[7] have been exchanged by the token “_MENTION_”.
5. All the HTML tags have been removed.
6. The hashtags[8] have been normalized with the token “_HASHTAG_”.
7. Tokens that express laughing (hahaha, hehehe...) have been normalized with the token “_LAUGH_”.
8. Users usually write expressions or abbreviations for surprise phrases like *omg*. All these kind of expressions are replaced by the token “_SURPRISE_”.
9. Positive emoticons like :), ;) or :, have been normalized with the token “_HAPPY_”.
10. Negative emoticons like :(, :’(or :-(have been normalized with the token “_SAD_”.
11. Twitter users usually repeat letters to emphasize the idea that they want to express. Therefore, all the words with a letter repeated more than two times have been reduced to only two instances. For example, the word “aaaamaaaaaziiiing” in tweet 111733236627025920 is transformed into “aamaaziing”.

After applying a normalization process to the training and development sets, we have used for the constrained system and the unconstrained system a dataset of 9,686 tweets.

3 Constrained System

The guidelines of the task define a constrained system as a system that only can use the train data provided by the organizers. Due to this restriction we decided to follow a supervised approach. It is required to define a set of parameters when the supervised method is the elected. The first step is to choose the minimum unit of information, i.e. what segments of text are considered as features. Pang et al. (2002) assert that

[7] A twitter mention is a reference to another user which has the pattern “*@user_name*”

[8] A hashtag is the way to refer a topic in Twitter, which has the pattern “*#topic_name*”

Class	Precision	Recall	F1-score
Positive	0.6983	0.6295	0.6621
Neutral	0.6591	0.8155	0.7290
Negative	0.5592	0.2710	0.3651
Average			0.6652

Table 1: Assessment with TF-IDF weighting scheme

Class	Precision	Recall	F1-score
positive	0.7037	0.6335	0.6668
neutral	0.6506	0.8313	0.7299
negative	0.5890	0.2105	0.3112
Average			0.6654

Table 2: Assessment with a binary weighting scheme

opinions or reviews should be represented with unigrams, but other work shows bigrams and trigrams outperformed the unigrams features (Dave et al., 2003). Therefore, there is not agreement in the SA research community about what is the best choice, unigrams or n-grams. Before several validations on the training set of the task we decided to use unigrams as feature for the polarity classification process. Thus, for the supervised algorithm, we have represented each tweet as a vector of unigrams.

The next decision was about the application of a stemmer process and getting rid off the English stop words. We only have applied stemmer process to the data because in previous works (Martínez-Cámara et al., 2013a) we did not reach good results removing the stop words in texts from Twitter. Another topic of discussion in the SA research community is the weighting scheme. Pang et al. (2002) weighted each unigram following a binary scheme. Also, in the most cited survey about SA (Pang and Lee, 2008) the authors indicated that the overall sentiment may not usually be highlighted through repeated use of the same terms. On the other hand, Martínez-Cámara et al. (2011) achieved the best results using TF-IDF as weighting scheme. Due to the lack of agreement in the SA research community about the use of a specific weight scheme, we have carried out several assessments with aim of deciding the most suitable one for the task. The machine learning algorithm selected for the evaluation was SVM. The results are shown in Tables 1 and 2.

The results achieved with the two weighting schemes are very similar. Regarding the positive class, the binary weighting scheme obtains better results than the TF-IDF one, so the presence of positive keywords is more useful than the frequent occurrence of those keywords. For the neutral class, regarding precision and F1-score, the TF-IDF scheme outperformed the binary scheme, but the recall had a higher value when the terms are weighted binary. The precision of the classification for the neutral class is only 1.2% better than the case where TF-IDF is used, while recall and the F1-score is better when the weighting of the features is binary. Although the negative class has a similar performance to that of the positive one with the two weighting schemes, we highlighted the high difference between the other two classes and the negative. The difference is more evident in the recall value, while the neutral class has a value of 0.8313 (binary), the negative one has a value of 0.2105 (binary). Therefore, due to the fact that the binary weighting scheme achieved better results in average, we decided to use it in the final system.

The last step in the configuration of a supervised approach based on machine learning is the selection of the algorithm. The algorithm selected was Support Vector Machine (SVM) (Cortes and Vapnik, 1995). Our decision is based on the widely used SVM by the research community of SA. The first application of SVM for SA was in (Pang et al., 2002) with good results. Since the publication of the previous work, other researchers have used SVM, and some of them are: (Zhang et al., 2009), (Pang and Lee, 2004) and (Jindal and Liu, 2006). Also, the algorithm SVM has been used to classify the polarity over tweets (Go et al., 2009) (Zhang et al., 2011) (Jiang et al., 2011). A broader review of the research about SA in Twitter can be found in (Martínez-Cámara et al., 2013b). Furthermore, our decision is supported by previous in-house experimentation.

For the experimentation we have used the framework for data mining RapidMiner[9]. In RapidMiner there are several implementations of SVM, among which we have selected LibSVM[10](Chang and Lin, 2011) with built-in default parametrization.

To sum up, the configuration of the SINAI constrained system is:

1. Machine learning approach: Supervised
2. Features: Unigrams.
3. Weighted scheme: Binary. If the term is presence the value is 1, 0 in other case.
4. Stemmer: Yes
5. Stopper: No
6. Algorithm: SVM.

The results reached during the development period are shown in Table 2

4 Unconstrained System

Our unconstrained system follows a two level categorization approach, determining whether the tweet is subjective or not at a first stage, and, for the subjective classified ones, whether the tweet is positive or negative. Both classification phases are fully based on knowledge resources. A predefined list of affective words is used for subjectivity detection, and a search process over the collection of emotions generated from a web resource is applied for final polarity classification. Figure 1 shows a general diagram of the system.

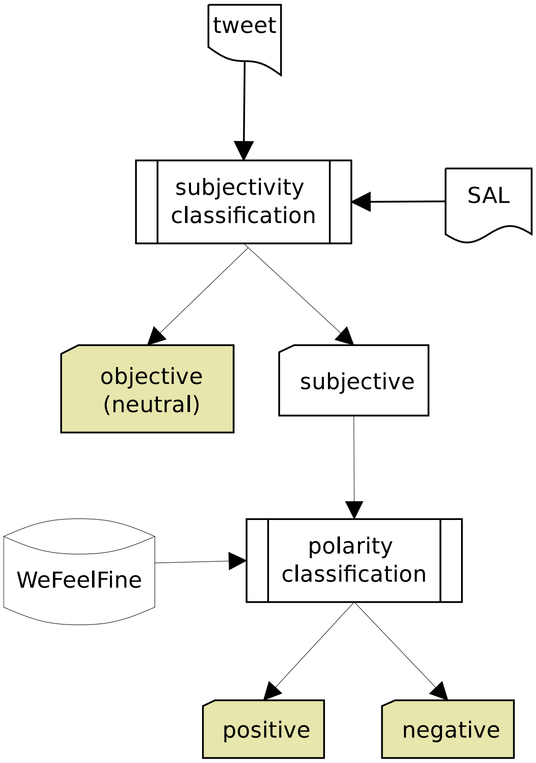

Figure 1: Unconstrained system general diagram

4.1 Step 1: determining subjectivity

The system based in WeFeelFine only categorizes between positive and negative texts, so a preliminary classification into subjective and objective (i.e. neutral) must be performed. To this end, a lexical approach is followed: those tweets containing at least one affective term from a list of predefined ones are considered subjective. If affective terms are not found, then the tweet is directly labeled as neutral. This list is called *Sentiment Analysis Lexicon* (SAL), which is defined in the work of Bing Liu (Hu and Liu, 2004). The list has two differentiated groups: a list of positive terms (*agile*, *enjoy*, *improving*) and another with negative ones (*anger*, *refusing*, *unable*...). At this phase, the polarity is not considered, so both lists are merged into a list of around 6,800 subjectivity terms.

4.2 Step 2: determining polarity

The WeFeelFine project (Kamvar and Harris, 2011) has been used as knowledge base for polarity classification following the approach proposed by (Montejo-Ráez, 2013). WeFeelFine[11] gathers affective texts from several blogs, creating a huge database of mood-related expressions. Almost two millions "feelings" are collected and indexed by the system. It is possible to retrieve related sentences and expressions by using its API. In this way, we have obtained the

[9] http://rapid-i.com/

[10] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[11] http://wefeelfine.org

top 200 most frequent feelings. For each feeling, about 1,500 sentences are include in a document that represents such a feeling. Then, using the Lucene[12] search engine, these documents have been indexed. In this way, we can use an incoming tweet as query and retrieve a ranked list of feelings, as shown in Figure 2.

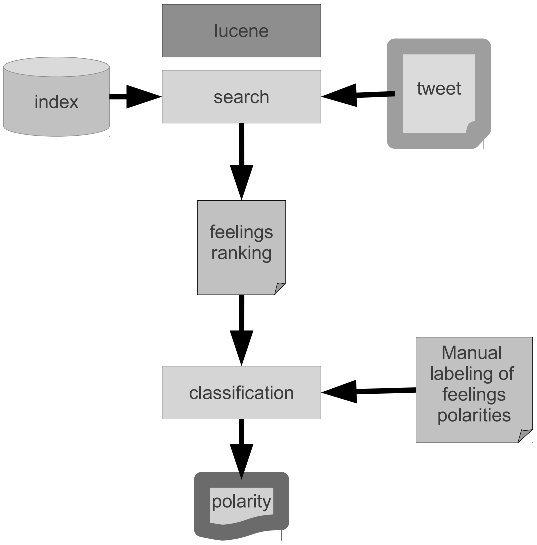

Figure 2: Polarity classification

The ranked list with the top 100 feelings (i.e. those feelings more related to the tweet) is taken for computing the final polarity by a summation of the manually assigned polarity of the feeling weighted with the score value returned by the engine, as shown in Equation 1.

$$p(t) = \frac{1}{|R|} \sum_{r \in R} RSV_r \cdot l_r \tag{1}$$

where
$p(t)$ is the polarity of tweet t
R is the list of retrieved feelings
l_r is the polarity label of feeling r
RSV_r is the *Ranking Status Value* of the feeling determined by Lucene.

As we did with the constrained system, we also assess the unconstrained system before applying the test data. The results reached during the evaluation phase are shown in Table 3. It is remarkable the fact that the precision value of the unconstrained system is a bit higher than the one reached by the constrained configuration. Thus, SAL is a good resource for subjective classification tasks. The unconstrained system reached worse results with positive and negative classes, but it is an expected result because supervised approaches usually obtain better results than the unsupervised and knowledge based approaches. However, the polarity classification has reached acceptable results, so it encourage us to follow improving the method based of the use of WeFeelFine.

Class	Precision	Recall	F1-score
positive	0.5004	0.6341	0.5593
neutral	0.6772	0.5416	0.6018
negative	0.3580	0.3456	0.3516
Average			0.5094

Table 3: Assessment of the unconstrained system

Acknowledgments

This work has been partially supported by a grant from the Fondo Europeo de Desarrollo Regional (FEDER), TEXT-COOL 2.0 project (TIN2009-13391-C04-02) and ATTOS project (TIN2012-38536-C03-0) from the Spanish Government. Also, this paper is partially funded by the European Commission under the Seventh (FP7 - 2007-2013) Framework Programme for Research and Technological Development through the FIRST project (FP7-287607). This publication reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

References

Enrique Amigó, Adolfo Corujo, Julio Gonzalo, Edgar Meij, and Md Rijke. 2012. Overview of replab 2012: Evaluating online reputation management systems. In *CLEF 2012 Labs and Workshop Notebook Papers*.

Chih-Chung Chang and Chih-Jen Lin. 2011. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27, May.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20:273–297.

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

Kushal Dave, Steve Lawrence, and David M. Pennock. 2003. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, WWW '03, pages 519–528, New York, NY, USA. ACM.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, pages 1–12.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 168–177, New York, NY, USA. ACM.

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 151–160, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nitin Jindal and Bing Liu. 2006. Identifying comparative sentences in text documents. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 244–251, New York, NY, USA. ACM.

Sepandar D. Kamvar and Jonathan Harris. 2011. We feel fine and searching the emotional web. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 117–126, New York, NY, USA. ACM.

Eugenio Martínez-Cámara, M. Teresa Martín-Valdivia, José M. Perea-Ortega, and L. Alfonso Ure na López. 2011. Opinion classification techniques applied to a spanish corpus. *Procesamiento de Lenguaje Natural*, 47.

Eugenio Martínez-Cámara, M. Teresa Martín-Valdivia, L. Alfonso Ure na López, and Ruslan Mitkov. 2013a. Detecting sentiment polarity in spanish tweets. *Information Systems Management*, In Press.

Eugenio Martínez-Cámara, M. Teresa Martín-Valdivia, L. Alfonso Ure na López, and Arturo Montejo-Ráez. 2013b. Sentiment analysis in twitter. *Natural Language Engineering*, FirstView:1–28, 2.

Arturo Montejo-Ráez. 2013. Wefeelfine as resource for unsupervised polarity classification. *Procesamiento del Lenguaje Natural*, 50:29–35.

Bo Pang and Lillian Lee. 2004. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, January.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA. Association for Computational Linguistics.

Julio Villena-Román, Sara Lana-Serrano, Eugenio Martínez-Cámara, and José Carlos González-Cristóbal. 2013. Tass - workshop on sentiment analysis at sepln. *Procesamiento del Lenguaje Natural*, 50(0).

Theresa Wilson, Zornitsa Kozareva, Preslav Nakov, Sara Rosenthal, Veselin Stoyanov, and Alan Ritter. 2013. SemEval-2013 task 2: Sentiment analysis in twitter. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '13, June.

Changli Zhang, Daniel Zeng, Jiexun Li, Fei-Yue Wang, and Wanli Zuo. 2009. Sentiment analysis of chinese documents: From sentence to document level. *Journal of the American Society for Information Science and Technology*, 60(12):2474–2487.

Ley Zhang, Riddhiman Ghosh, Mohamed Dekhil, Meichun Hsu, and Bing Liu. 2011. Combining lexiconbased and learning-based methods for twitter sentiment analysis. *HP Laboratories, Technical Report HPL-2011-89*.