Supervised Polarity Classification of Spanish Tweets based on Linguistic Knowledge

David Vilares
Universidade da Coruña
Campus de Elviña, 15071
A Coruña, Spain
david.vilares@udc.es

Miguel A. Alonso
Universidade da Coruña
Campus de Elviña, 15071
A Coruña, Spain
miguel.alonso@udc.es

Carlos Gómez-Rodríguez
Universidade da Coruña
Campus de Elviña, 15071
A Coruña, Spain
carlos.gomez@udc.es

ABSTRACT

We describe a system that classifies the polarity of Spanish tweets. We adopt a hybrid approach, which combines machine learning and linguistic knowledge acquired by means of NLP. We use part-of-speech tags, syntactic dependencies and semantic knowledge as features for a supervised classifier. Lexical particularities of the language used in Twitter are taken into account in a pre-processing step. Experimental results improve over those of pure machine learning approaches and confirm the practical utility of the proposal.

Categories and Subject Descriptors

H.3.1 [**Information Retrieval and Storage**]: Content Analysis and Indexing—*Linguistic processing*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Text analysis*

Keywords

Document Analysis, Linguistic Analysis, Machine Learning, Opinion Mining, Sentiment Analysis, Twitter

1. INTRODUCTION

Opinion Mining (OM) has become a relevant field of research in the last decade. With the explosion of the Web 2.0, many users employ social media to share their opinions and experiences about products, services or relevant people. In this context, one of the most popular social media is Twitter. In this microblogging network, users express their views in micro documents (tweets) of up to 140 characters, particularly about current topics, which is an important source of information for companies, especially for their business intelligence and marketing departments.

We present a system which classifies the polarity of Spanish tweets taking into account linguistic knowledge. We use as our starting point an external semantic-based OM system, using its output as features for a supervised classifier. We then include POS-tags and syntactic information in the form of syntactic dependencies. Lastly, we provide an automatic mechanism to enrich and adapt semantic knowledge to a specific domain. We evaluate our proposal with the TASS 2012 corpus, which distinguishes between six different categories. The remainder of this paper is organized as follows. We start by describing our proposal in Section 2. In Section 3 we show the experimental results. Finally, we present conclusions and future work in Section 4.

DocEng'13, September 10–13, 2013, Florence, Italy.

ACM 978-1-4503-1770-2/13/09...$15.00.
http://dx.doi.org/10.1145/2494266.2494300

2. POLARITY CLASSIFICATION HYBRID SYSTEM

The polarity classification task has mainly been tackled from two different perspectives: semantic-based [15] and supervised [9]. Semantic approaches are characterised by the use of semantic orientation (SO) dictionaries or opinion lexicons. They have been applied successfully in many contexts, but their performance drops on Twitter, where there is a high frequency of subjective elements, such as emoticons or Twitter special expressions, which are not included in general opinion lexicons; this results in a low recall [18].

With respect to supervised classifiers, their main drawback is their high domain dependency and the cost of creating new training data. Supervised methods typically represent the text as a bag of words, learning the perception of a word for a specific context. However, their performance drops drastically when the same classifier is used to categorise texts from a different field [13]. In contrast with these approaches, we propose a hybrid system which combines lexical, syntactic and semantic knowledge with machine learning techniques. In particular, linguistic features are used to feed an SMO, an implementation of SVM, presented in [10], and incorporated by default in the WEKA data mining software [5]. Figure 1 shows the general architecture of our proposal, whose components will be described in Section 2.

To train and evaluate our approach we use the TASS 2012 corpus, presented at the Workshop on Sentiment Analysis at SEPLN[1] [17]. It is a collection of Spanish tweets written by public figures that is composed of a training and a test set which contain 7,219 and 60,798 tweets, respectively. Each one is annotated with one of these six categories: *strongly positive* (P+), *positive* (P), *neutral* (NEU), *negative* (N), *strongly negative* (N+) or *without opinion* (NONE). An annotation in four classes was also proposed (P+ and N+ classes are included into P and N, respectively). The gold standard has been generated by a pooling of the submis-

[1]Sociedad Española para el Procesamiento del Lenguaje Natural

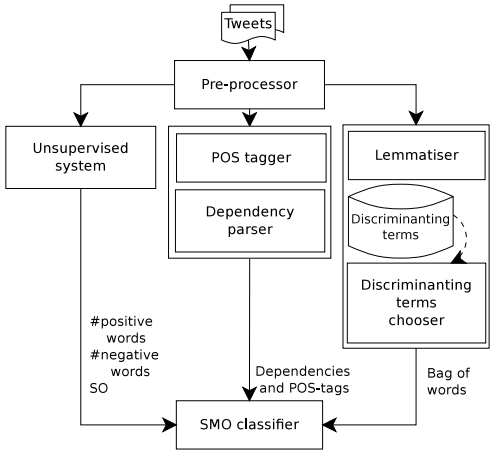

Figure 1: General architecture of the system

sions of the workshop followed by a thorough human review for the thousands of ambiguous cases.

2.1 Pre-processing

As a previous step, all tweets were pre-processed as follows:

- *Emoticon replacement*: We employ the emoticon collection published in [1]. Each emoticon is replaced in the text by one of these five labels: strong positive (ESP), positive (EP), neutral (ENEU), negative (EN) or strong negative (ESN).
- *Most frequent unrecognised abbreviations spell-checking*: We replace some of the most habitual ungrammatical Spanish abbreviations by their grammatical form.
- *URL normalisation*: Web addresses are replaced with the string *'URL'*.
- *Laughs normalisation*: Different variants of laughs in Spanish language (*e.g. 'jjjaja'*, 'JJEEJJ',...) are normalised as *jxjx* where x $\in \{a, e, i, o, u\}$. For example, 'jjjaja' becomes 'jaja'.
- *Treatment of special Twitter elements ('@'and '#')*: User mentions are modified: we eliminate the '@' symbol and capitalise the first character (*e.g. '@user' becomes 'User'*). Regarding the hashtags, if one appears at the beginning or the end of a tweet, then the complete hashtag is eliminated. Otherwise we only delete the '#'.

2.2 Generic semantic approach (GSA)

We took a purely semantic proposal presented in [16] as a part of our system. This approach carries out segmentation, tokenisation and POS-tagging of texts, to then obtain a dependency tree for each sentence by means of dependency parsing. We then employ the syntactic structure to treat three of the most significant constructions on sentiment analysis: intensification, adversative subordinate clauses and negation. We use dependency types to identify the scope of these constructions and modify the semantic orientation value of the polarity words inside that scope. As a result, we obtain the SO for each sentence and then aggregate them to calculate the global SO of a text.
We finally include that final SO, and the number of positive and negative words in a tweet, as features for a supervised classifier. We use the Spanish opinion lexicon of Brooke et al. [3] to determine which words are opinionated.

2.3 Morphosyntactic information (MSI)

The employment of POS-tagging information in polarity classification tasks is a widely discussed issue. Pak and Paroubek [8] and Spencer and Uchyigit [12] suggest that certain POS-tags, such as adjectives or personal pronouns, are more frequent in subjective texts. In this respect, we observed a similar tendency in the training set of the TASS 2012 corpus. Table 1 shows a selection of relevant tag frequencies. In the same way, we hypothesise that dependency types are also useful in order to classify the polarity of the tweets.

To test this, we have used the Ancora corpus [14] and the Nivre arc-eager algorithm included in MaltParser [6].

Class	a	n	v	i	f
P+	0.060	0.256	0.111	0.004	0.215
P	0.056	0.266	0.119	0.002	0.198
NEU	0.057	0.254	0.133	0.001	0.163
N	0.050	0.263	0.132	0.001	0.161
N+	0.060	0.266	0.118	0.001	0.154
NONE	0.048	0.299	0.090	0.001	0.220

Table 1: Tag frequencies in the training set: adjectives (*a*), nouns (*n*), verbs (*v*), interjections (*i*) and punctuation marks (*f*)

Class	ci	atr	cc	cag
P+	0.008	0.105	0.042	0.004
P	0.010	0.010	0.051	0.000
NEU	0.010	0.141	0.053	0.001
N	0.009	0.000	0.055	0.150
N+	0.007	0.000	0.049	0.145
NONE	0.179	0.008	0.003	0.001

Table 2: Dependency type frequencies in the training set: indirect object (*ci*), subject complement (*atr*), adjunct (*cc*) and agent (*cag*)

Table 2 shows the frequency of some dependency types[2] on the training set of the TASS 2012 corpus. The frequency distribution of certain dependencies such as the agent is especially relevant, which suggests that Spanish users employ the passive voice more often in negative reviews. To treat both POS-tags and dependency structure, we included the total number of occurrences of each tag and dependency type instance found in each tweet as features for the classifier.

2.4 Domain adaptation (DA)

In Twitter texts, there is a high frequency of some special subjective elements that are not included in generic opinion lexicons. Emoticons, laughs and some Twitter tags, such as *Follow Friday* (*'FF'*) or *Retweet* (*'RT'*), are some of the

[2]We use the Ancora dependency type tags.

clearest examples that, while usually being subjective, do not appear in semantic dictionaries. To improve the performance in a specific domain and in a specific social medium, we have developed an automatic mechanism that enriches and adapts semantic knowledge to a particular field. Our procedure consists of two different and separate tasks: *selection of the most discriminating tokens* and *adaptation of semantic dictionaries*.

2.4.1 Selection of the most discriminating tokens

The goal is to create a ranked list of words to help distinguish between the different categories of the TASS 2012 corpus, and use each word of that list as a feature for the classifier. We use binary occurrence as the weighting factor, because we hypothesise that each word usually appears at most once in a tweet.

Term	Ranking (4 classes)	Ranking (6 classes)
EP (emoticon)	1	1
URL	4	4
FF	30	47
jaja (laugh)	101	11,964

Table 3: Ranking of some of discriminating terms on the training set of the TASS 2012 corpus

We rank the terms by measuring the *information gain* with respect to the class, employing the attribute selection tools provided by WEKA and the training set of the TASS 2012 corpus. To make the selection more robust we used a ten-fold cross-validation. We extracted more than 14,000 discriminating terms. However, only a few hundred of terms provided an information gain greater than zero, so we decide to include only those words. Table 3 shows some effective classifier tokens that are not included in a generic dictionary.

2.4.2 Adaptation of semantic dictionaries

Our generic semantic approach uses a generic opinion lexicon. In order to adapt it to the Twitter domain, we have developed an automatic enrichment mechanism for semantic dictionaries. In the same line as in *Selection of the most discriminating tokens*, we rank the best polarity terms, but in this case, we have only taken into account P+, P, N and N+ classes. We then assign a SO to each ranked term, based on the number of occurrences both in positive and negative tweets, and we add them to the semantic dictionaries. However, the improvement in performance was negligible when this method was used jointly with the *Selection of the most discriminating tokens*.

3. EXPERIMENTAL RESULTS

The TASS 2012 workshop proposed two different tasks about sentiment analysis: classification into six categories (P+, P, NEU, N, N+ and NONE) and classification into four categories (the classes P+ and N+ are included in the classes P and N, respectively). We used the TASS 2012 training and test sets to evaluate our proposal.

Tables 4 and 5 show the results obtained for the two polarity classification tasks: four and six categories. We used the F-measure defined as $F = \frac{2 \times R \times P}{R+P}$, where P is the number of true positives divided by the sum of true and false positives, and R is the number of true positives divided by the sum of the true positives and false negatives. The subscripts in Tables 4 and 5 refer to each category of the TASS 2012 corpus. In both tasks, the GSA approach obtains a good performance. The incorporation of POS-tag and syntactic information improves the classification performance on positive and negative tweets. This reinforces the idea that users employ certain POS-tags and syntactic functions more frequently depending on the polarity of the review. The accuracy obtained by our final approach suggests that, although generic opinion lexicons and the morphosyntactic structure of the tweets are helpful to classify the sentiment of the message we need to incorporate domain semantic knowledge to optimise the performance.

Measure	GSA	MSI	DA
F_p	0.631	0.680	0.745
F_{neu}	0.000	0.000	0.054
F_n	0.566	0.603	0.671
F_{none}	0.574	0.564	0.620
Accuracy	0.587	0.615	0.676

Table 4: Results on the test set (4 classes)[2]

Measure	GSA	MSI	DA
F_{p+}	0.609	0.637	0.705
F_p	0.000	0.040	0.307
F_{neu}	0.000	0.009	0.089
F_n	0.452	0.478	0.512
F_{n+}	0.000	0.120	0.441
F_{none}	0.575	0.605	0.648
Accuracy	0.523	0.546	0.600

Table 5: Results on the test set (6 classes)

Moreover, in both cases, the performance on neutral tweets is low.[3] We hypothesise that this phenomenon is due to two factors. The first refers to an intrinsic characteristic of neutral tweets: the mixture of favourable and unfavourable opinions complicates the categorisation of these tweets, even more so in Twitter, where users have no space to argue their point of view. The second refers to the ambiguous criteria used in the corpus to distinguish between NEU and NONE tweets, as has been pointed out by some authors [11].

Method	Accuracy (4 classes)	Accuracy (6 classes)
Our proposal (features+SMO)	0.676	0.600
SMO	0.630	0.532
Our features with NaiveBayes	0.582	0.494
Our features with j48	0.574	0.452
j48	0.565	0.482
NaiveBayes	0.523	0.472

Table 6: Performance on the TASS 2012 test set with different methods

Finally, we tested the effectiveness of our features with other classifiers. We selected NaiveBayes and j48 (the WEKA implementation of a C4.5 decision tree). Table 6 shows the per-

[3]The small number of the NEU tweets into the TASS 2012 training set (around 2%) makes it difficult for the classifier to learn these tweets satisfactorily.

formance of these classifiers, compared to the corresponding pure machine learning approaches, which use as attributes a vector of words representing the text. In this case, we applied pre-processing of Section 2.1 and lemmatisation steps and we kept the WEKA default configuration. Results suggest that our features are generalizable and outperform the baseline of different classifiers.

4. CONCLUSIONS AND FUTURE WORK

In this paper we describe an approach which uses POS-tag information, dependency structure and semantic knowledge to train a supervised classifier that categorises the sentiment of Spanish tweets. Experimental results show a good performance and suggest that the morphosyntactic structure of the tweets is useful to classify their sentiment.

As future work, there are many aspects that we would like to explore. The current preprocessing of the tweets is quite simple. We would like to determine how an exhaustive normalisation of tweets could help in polarity classification tasks. In this respect, Oliva et al. [7] propose a SMS normalisation system that could enrich our preprocessing module. We would also like to explore how to modify dependency parsing for microtexts. In this line, the approach of Gimpel et al. [4] could be usefully adapted to our proposal. Finally, we believe that the method of Batista and Ribeiro [2] could help to improve the performance of our approach: instead of training a classifier to distinguish between n categories, they train *n-1* binary classifiers and combine the final results, exploiting the differences between the different classes.

Acknowledgments

Research reported in this paper has been partially funded by Ministerio de Economía y Competitividad and FEDER (Grant TIN2010-18552-C03-02) and by Xunta de Galicia (Grants CN2012/008, CN2012/319).

5. REFERENCES

[1] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau. Sentiment analysis of Twitter data. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, pages 30–38, Stroudsburg, PA, USA, 2011. ACL.

[2] F. Batista and R. Ribeiro. The L2F Strategy for Sentiment Analysis and Topic Classification. *Procesamiento de Lenguaje Natural*, 50:77–84, 2013.

[3] J. Brooke, M. Tofiloski, and M. Taboada. Cross-Linguistic Sentiment Analysis: From English to Spanish. In *Proceedings of the International Conference RANLP-2009*, pages 50–54, Borovets, Bulgaria, 2009. ACL.

[4] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. Smith. Part-of-speech tagging for Twitter: annotation, features, and experiments. *HLT '11 Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers*, 2:42–47, 2011.

[5] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten. The WEKA data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, Nov. 2009.

[6] J. Nivre, J. Hall, J. Nilsson, A. Chanev, G. Eryigit, S. Kübler, S. Marinov, and E. Marsi. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135, 2007.

[7] J. Oliva, J. I. Serrano, M. D. Del Castillo, and A. Igesias. A SMS normalization system integrating multiple grammatical resources. *Natural Language Engineering*, 19:121–141, 2013.

[8] A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh International conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA).

[9] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, pages 79–86, 2002.

[10] J. C. Platt. Advances in kernel methods. chapter Fast training of support vector machines using sequential minimal optimization, pages 185–208. MIT Press, Cambridge, MA, USA, 1999.

[11] X. Saralegi Urizar and I. San Vicente Roncal. Detecting Sentiments in Spanish Tweets. In *TASS 2012 Working Notes*, CastellÃşn, Spain, 2012.

[12] J. Spencer and G. Uchyigit. Sentimentor: Sentiment Analysis on Twitter Data. In *The 1st International Workshop on Sentiment Discovery from Affective Data*, Bristol, United Kingdom, 2012.

[13] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307, 2011.

[14] M. Taulé, M. A. Martí, and M. Recasens. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, and D. Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, 2008.

[15] P. D. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424, Stroudsburg, PA, USA, 2002. ACL.

[16] D. Vilares, M. A. Alonso, and C. Gómez-Rodríguez. Clasificación de polaridad en textos con opiniones en español mediante análisis sintáctico de dependencias. *Procesamiento de Lenguaje Natural*, 50:13–20, 2013.

[17] J. Villena-Román, S. Lana-Serrano, J. C. González Cristóbal, and E. Martínez-Cámara. TASS - Worshop on Sentiment Analysis at SEPLN. *Procesamiento de Lenguaje Natural*, 50:37–44, 2013.

[18] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B. Liu. Combining lexicon-based and learning-based methods for Twitter sentiment analysis. Technical Report HPL-2011-89, HP Laboratories, Palo Alto, CA, 2011.