# LyS: Porting a Twitter Sentiment Analysis Approach from Spanish to English

**David Vilares, Miguel Hermo, Miguel A. Alonso, Carlos Gómez-Rodríguez, Yerai Doval**
Grupo LyS, Departamento de Computación, Facultade de Informática
Universidade da Coruña, Campus de A Coruña
15071 A Coruña, Spain
{david.vilares, miguel.hermo, miguel.alonso, carlos.gomez, yerai.doval}@udc.es

## Abstract

This paper proposes an approach to solve message- and phrase-level polarity classification in Twitter, derived from an existing system designed for Spanish. As a first step, an *ad-hoc* preprocessing is performed. We then identify lexical, psychological and semantic features in order to capture different dimensions of the human language which are helpful to detect sentiment. These features are used to feed a supervised classifier after applying an information gain filter, to discriminate irrelevant features. The system is evaluated on the SemEval 2014 task 9: Sentiment Analysis in Twitter. Our approach worked competitively both in message- and phrase-level tasks. The results confirm the robustness of the approach, which performed well on different domains involving short informal texts.

## 1 Introduction

Millions of opinions, conversations or just trivia are published each day in Twitter by users of different cultures, countries and ages. This provides an effective way to poll how people praise, complain or discuss about virtually any topic. Comprehending and analysing all this information has become a new challenge for organisations and companies, which aim to find out a way to make quick and more effective decisions for their business. In particular, identifying the perception of the public with respect to an event, a service or an entity are some of their main goals in a short term. In this respect, *sentiment analysis*, and more specifically *polarity classification*, is playing an important role in order to automatically analyse subjective information in texts.

This paper describes our participation at SemEval 2014 task 9: Sentiment Analysis in Twitter. Specifically, two subtasks were presented: (A) contextual polarity disambiguation and (B) message polarity classification. The first subtask consists on determining the polarity of words or phrases extracted from short informal texts, the scope of extracts being provided by the SemEval organisation. Subtask B focusses on classifying the content of the whole message. In both cases, three possible sentiments are considered: *positive*, *negative* and *neutral* (which involves mixed and non-opinionated instances). Although the training set only contains tweets, the test set also includes short informal texts from other domains, in order to measure cross-domain portability. You can test the model for subtask B at `miopia.grupolys.org`.

## 2 SemEval 2014-Task 9: Sentiment Analysis in Twitter

Our contribution is a reduced version of a Spanish sentiment classification system (Vilares et al., 2013a; Vilares et al., 2013b) that participated in TASS 2013 (Villena-Román et al., 2014), achieving the 5th place on the global sentiment classification task and the 1st place on topic classification on tweets. In this section we describe how we have ported to English this system originally designed for Spanish. Tasks A and B are addressed from the same perspective, which is described below.

### 2.1 Preprocessing

We implement a naive preprocessing algorithm which seeks to normalise some of the most common ungrammatical elements. It is intended for Twitter, but many of the issues addressed would also be valid in other domains:

- *Replacement of frequent abbreviations* The list of the most frequent ones was extracted from the training set, taking the Penn Treebank (Marcus et al., 1993) as our dictionary. A term is considered ungrammatical if it does not appear in our dictionary. We then carry out a manual review to distinguish between unknown words and abbreviations, providing a correction in the latter case. For example, *'c'mon'* becomes *'come on'* and *'Sat'* is replaced by *'Saturday'*.

- *Emoticon normalisation*: We employ the emoticon collection published in (Agarwal et al., 2011). Each emoticon is replaced with one of these five labels: strong positive (ESP), positive (EP), neutral (ENEU), negative (EN) or strong negative (ESN).

- *Laughs* : Multiple forms used in social media to reflect laughs (*e.g. 'hhahahha'*, *'HHEHEHEH'*) are preprocessed in a homogeneous way to obtain a pattern of the form *'hxhx'* where $x \in \{a, e, i, o, u\}$.

- *URL normalisation*: External links are replaced by the string *'url'*.

- *Hashtags ('#') and usernames ('@')*: If the hashtag appears at the end or beginning of the tweet, we remove the hashtag. Based on other participant approaches at SemEval 2013 (Nakov et al., 2013), we realized maybe this is not the best option, although we believe hashtags will not be useful in most of cases, since they refer to very specific events. Otherwise, only the '#' is removed, hypothesising the hashtag is used to emphasise a term (e.g. *'Matthew #Mcconaughey has won the Oscar'*).

## 2.2 Feature extraction

Our approach only takes into account information extracted from the text, without considering any kind of meta-data. Extracted features combine lexical, psychological and semantic knowledge in order to build a linguistic model able to analyse tweets, but also other kinds of messages. These features can be divided into two types: *corpus-extracted features* and *lexicon-extracted features*. All of them take the total number of occurrences of the respective feature as the weighting factor to then feed the supervised classifier.

### 2.2.1 Corpus-extracted features

Given a corpus, we use it to extract the following set of features:

- *Word forms*: A model based on this type of features is our baseline. Each single word is considered as a feature in order to feed the supervised classifier. This often becomes a simple and acceptable start point which obtains a decent performance.

- *Part-of-speech (PoS) information*: some coarse-grained PoS-tags such as *adjective* or *adverb* are usually good indicators of subjective texts while some fine-grained PoS tags such as *third person personal pronoun* provide evidence of non-opinionated messages (Pak and Paroubek, 2010).

### 2.2.2 Lexicon-extracted features

We also consider information obtained from external lexicons in order to capture linguistic information that can not be extracted from a training corpus by means of bag-of-words and PoS-tag models. We rely on two manually-build lexicons:

- *Pennebaker et al. (2001) psychometric dictionaries*. Linguistic Inquiry and Word Count[1] (LIWC) is a software which includes a semantic dictionary to measure how people use different kinds of words over a wide number of texts. It categorises terms into *psychometric properties*, which correspond to different dimensions of the human language. The dictionary relates terms with psychological properties (*e.g. anger* or *anxiety*), but also with topics (*e.g. family*, *friends*, *religion*) or even morphological features (*e.g. future time*, *past time* or *exclamations*).

- *Hu and Liu (2004) opinion lexicon*. It is a collection of positive and negative words. Many of the occurrences are misspelled, since they often come from web environments.

### 2.2.3 Syntactic features

We also parsed the tweets using MaltParser (Nivre et al., 2007) in order to obtain dependency triplets of the form $(w_i,\ arc_{ij},\ w_j)$, where $w_i$ is the head word $w_j$, the dependent one and $arc_{ij}$ the existing syntactic relation between them. We tried to incorporate generalised dependency triplets (Joshi

[1] http://www.liwc.net/

and Penstein-Rosé, 2009), following an enriched perspective presented in Vilares et al. (2014a). A generalisation consists on backing off the words to more abstracted terms. For example, a valid dependency triplet for the phrase *'awesome villain'* is *(villain, modifier, awesome)*, which could be generalised into *(anger, modifier, assent)* by means of psychometric properties. However, experimental results over the development corpus using these features decreased performance with respect to our best model, probably due to the small size of the training corpus, since dependency triplets tend to suffer from sparsity, so a larger training corpus is needed to exploit them in a proper way (Vilares et al., 2014a).

### 2.3 Feature selection

For a machine learning approach, sparsity could be an issue. In particular, due to the size of the corpus, many of the terms extracted from the training set only appear a few times in it. This makes it impossible to properly learn the polarity of many tokens. Thus, we carry out a filtering step before feeding our classifier. In particular, we rely on the information gain (IG) method to then rank the most relevant features. Information gain measures the relevance of an attribute with respect to a class. It takes values between 0 and 1, where a higher value implies a higher relevance. Table 1 shows the top five relevant features based on their information gain for our best model. The top features for task A were very similar. Our official runs only consider features with an IG greater than zero.

| IG | Feature | Category |
|---|---|---|
| 0.140 | positive emotion | Pennebaker et al. (2001) |
| 0.137 | #positive-words | Hu and Liu (2004) |
| 0.126 | affect | Pennebaker et al. (2001) |
| 0.089 | #negative-words | Hu and Liu (2004) |
| 0.083 | negative emotion | Pennebaker et al. (2001) |

Table 1: Most relevant features for task B. '#' must be read this table as 'the number of'and not as a hashtag.

### 2.4 Classifier

We have trained our runs with a SVM LibLINEAR classifier (Fan et al., 2008) taking the implementation provided in WEKA (Hall et al., 2009). The selection was motivated by the acceptable results that some of the participants in SemEval 2013, e.g. Becker et al. (2013), obtained using this implementation. We configured the multi-class support vector machine by Crammer and Singer (2002) as the SVMtype. Since the corpus was unbalanced, we tuned the weights for the classes using the development corpus: 1 for the *positive* class, 2 for *negative* and 0.5 for *neutral*. The rest of parameters were set to default values.

## 3 Experimental results

The SemEval 2014 organisation provides a standard training corpus for both tasks A an B. For task A, each tweet is marked with a list of the words and phrases to analyse, and for each one its sentiment label is provided. In addition, a development corpus was released for tuning the system parameters. The training and the development corpus can be used jointly (*constrained runs*) to train models that are then evaluated over the test corpus.[2] Some participants used external annotated corpora (*unconstrained runs*) to build their models. With respect to the test corpus, it contains texts from tweets but also from LiveJournal texts, which we are abbreviating as LJ, and SMS messages.

Table 2 contains the statistics of the corpora we used. Sharing data is a violation of Twitter's terms of service, so we had to download them. Unfortunately, some of the tweets were no longer available for several reasons, e.g., user or a tweet does not exist anymore or the privacy settings of a user have changed. As a result, the size of our training and development corpora may be different from those of other participant's corpora.

| Task | Set | Positive | Negative | Neutral |
|---|---|---|---|---|
| A | Train | 4,917 | 2,591 | 385 |
| | Dev | 555 | 365 | 45 |
| | Test | 6,354 | 3,771 | 556 |
| B | Train | 3,063 | 1,202 | 3,935 |
| | Dev | 493 | 290 | 633 |
| | Test | 3,506 | 1,541 | 3,940 |

Table 2: SemEval 2014 corpus statistics

### 3.1 Evaluation metrics

F-measure is the official score to measure how systems behave on each class. In order to rank participants, the SemEval 2014 organisation proposed the averaged F-measure of positive and negative tweets.

[2] We followed this angle.

### 3.2 Performance on sets

Tables 3 and 4 show performance on the test set of different combinations of the proposed features. Table 5 shows the performance of our run on task A. The results over the corresponding sets for task B are illustrated in Table 6. They are significant lower than in task A. This suggests that when a message involves more than one of two tokens, a lexical approach is not enough. Improving performance should involve taking into account context and linguistic phenomena that appear in sentences to build a model based on the composition of linguistic information.

| Model | LJ | SMS | Twitter 2013 | Twitter 2014 | Twitter Sarcasm |
|---|---|---|---|---|---|
| WPLT (no IG) | 82.21 | **82.32** | 84.82 | **81.69** | 71.19 |
| WPL | 83.55 | 81.04 | 84.85 | 80.64 | 68.79 |
| WPLT* | **83.96** | 81.46 | **85.63** | 79.93 | 71.98 |
| WP | 78.53 | 80.97 | 80.34 | 73.35 | **74.18** |
| P | 75.70 | 78.74 | 73.58 | 65.75 | 71.82 |
| W | 61.58 | 65.45 | 64.56 | 59.16 | 62.93 |
| L | 66.04 | 64.11 | 62.96 | 53.81 | 61.26 |
| T | 47.07 | 51.37 | 71.82 | 43.64 | 49.37 |

Table 3: Performance on the test set for task A. The model marked with a * was our official run. W stands for features obtained from a bag-of-words approach, L from Hu and Liu (2004), P from Pennebaker et al. (2001) and T for fine-grained PoS-tags. They can be combined, e.g., a model named WP use both words and psychometric properties.

| Model | LJ | SMS | Twitter 2013 | Twitter 2014 | Twitter Sarcasm |
|---|---|---|---|---|---|
| WPLT* | 69.79 | 60.45 | **66.92** | **64.92** | 42.40 |
| WPL | **70.19** | **61.41** | 66.71 | 64.51 | 45.72 |
| WP | 66.84 | 60.22 | 65.29 | 63.90 | 45.90 |
| WPLT (no IG) | 66.38 | 57.01 | 61.96 | 62.84 | 43.71 |
| W | 65.12 | 56.00 | 62.87 | 62.64 | 48.75 |
| P | 63.42 | 54.80 | 60.05 | 57.66 | **54.20** |
| T | 45.99 | 35.85 | 46.53 | 45.99 | 48.58 |
| L | 57.53 | 45.14 | 48.80 | 44.48 | 49.14 |

Table 4: Performance on the test set for task B

## 4 Conclusions

This papers describes the participation of the LyS Research Group (`http://www.grupolys.org`) at the SemEval 2014 task 9: Sentiment Analysis in Twitter, with a system that attained competitive performance both in message and phrase-level tasks, as can be observed in Table 7. This system is a reduced version of a sentiment classification model for Spanish texts that performed well

| Test set | Positive | Negative | Neutral |
|---|---|---|---|
| DEV | 86.30 | 81.60 | 4.30 |
| TWITTER 2013 (full) | 88.70 | 81.90 | 17.60 |
| TWITTER 2013 (progress subset) | 88.81 | 82.57 | 20.75 |
| LJ | 84.34 | 83.56 | 13.84 |
| SMS | 80.31 | 82.56 | 7.10 |
| TWITTER 2014 | 89.02 | 70.82 | 4.44 |
| TWITTER SARCASM | 85.71 | 57.63 | 28.57 |

Table 5: Performance on different sets for our model on task A. The model evaluated on the development set was only built using the training set.

| Test set | Positive | Negative | Neutral |
|---|---|---|---|
| DEV | 69.80 | 60.40 | 66.70 |
| TWITTER 2013 (full) | 72.50 | 64.30 | 72.30 |
| TWITTER 2013 (progress subset) | 71.92 | 61.92 | 71.22 |
| LJ | 71.94 | 67.65 | 66.23 |
| SMS | 63.83 | 57.06 | 73.76 |
| TWITTER 2014 | 74.26 | 55.58 | 66.76 |
| TWITTER SARCASM | 55.17 | 29.63 | 51.61 |

Table 6: Performance on different sets for our model on task B

| Test set | Task A | Task B |
|---|---|---|
| LiveJournal 2014 | 4 / 27 | 13 / 50 |
| SMS 2013 | 12 / 27 | 19 / 50 |
| Twitter 2013 | 9 / 27 | 10 / 50 |
| Twitter 2014 | 11 / 27 | 18 / 50 |
| Twitter 2014 Sarcasm | 10 / 27 | 33 / 50 |

Table 7: Position of our submission on each corpus and task, according to results provided by the organization on April 22, 2014

in the TASS 2013 (Villena et al., 2013). The official results show how our approach works competitively both on tasks A and B without needing large and automatically-built resources. The approach is based on a bag-of-words that includes word-forms and PoS-tags. We also extract psychometric and sentiment information from external lexicons. In order to reduce sparsity problems, we firstly apply an information gain filter to select only the most relevant features. Experiments on the development set showed a significant improvement on the same model with respect to skipping it on subtask B.

As future work, we plan to incorporate syntactic information as we did for the original system for Spanish language (Vilares et al., 2014b).

## Acknowledgements

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 (Grant CN2012/008).

## References

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

Lee Becker, George Erhart, David Skiba, and Valentine Matula. 2013. Avaya: Sentiment analysis on twitter with self-training and polarity lexicon expansion. *Atlanta, Georgia, USA*, page 333.

Koby Crammer and Yoram Singer. 2002. On the algorithmic implementation of multiclass kernel-based vector machines. *The Journal of Machine Learning Research*, 2:265–292.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. 2009. The weka data mining software: an update. *SIGKDD Explorations*, 11(1):10–18, November.

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*, pages 168–177. ACM.

M. Joshi and C. Penstein-Rosé. 2009. Generalizing dependency features for opinion mining. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, ACLShort '09, pages 313–316, Suntec, Singapore. Association for Computational Linguistics.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

Preslav Nakov, Zornitsa Kozareva, Alan Ritter, Sara Rosenthal, Veselin Stoyanov, and Theresa Wilson. 2013. Semeval-2013 task 2: Sentiment analysis in twitter.

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

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

J.W. Pennebaker, M.E. Francis, and R.J. Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, page 71.

David Vilares, Miguel A. Alonso, and Carlos Gómez-Rodríguez. 2013a. LyS at TASS 2013: Analysing Spanish tweets by means of dependency parsing, semantic-oriented lexicons and psychometric word-properties. In Alberto Díaz Esteban, Iñaki Alegría Loinaz, and Julio Villena Román, editors, *XXIX Congreso de la Sociedad Española de Procesamiento de Lenguaje Natural (SEPLN 2013). TASS 2013 - Workshop on Sentiment Analysis at SEPLN 2013*, pages 179–186, Madrid, Spain, September.

David Vilares, Miguel A. Alonso, and Carlos Gómez-Rodríguez. 2013b. Supervised polarity classification of Spanish tweets based on linguistic knowledge. In *DocEng'13. Proceedings of the 13th ACM Symposium on Document Engineering*, pages 169–172, Florence, Italy, September. ACM.

David Vilares, Miguel A. Alonso, and Carlos Gómez-Rodríguez. 2014a. On the usefulness of lexical and syntactic processing in polarity classification of twitter messages. *Journal of the Association for Information Science Science and Technology*, to appear.

David Vilares, Miguel A. Alonso, and Carlos Gómez-Rodríguez. 2014b. A syntactic approach for opinion mining on spanish reviews. *Natural Language Engineering*. Available on CJO2013. doi:10.1017/S1351324913000181.

Julio Villena-Román, Janine García-Morera, Cristina Moreno-García, Sara Lana-Serrano, and José Carlos González-Cristóba. 2014. TASS 2013 — a second step in reputation analysis in Spanish. *Procesamiento del Lenguaje Natural*, 52:37–44, March.