# LyS at TASS 2015: Deep Learning Experiments for Sentiment Analysis on Spanish Tweets[*]

## *LyS en TASS 2015: Experimentos con Deep Learning para Análisis del Sentimiento sobre Tweets en Español*

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

**Resumen:** Este artículo describe la participación del grupo LyS en el TASS 2015. En la edición de este año, hemos utilizado una red neuronal denominada *long short-term memory* para abordar los dos retos propuestos: (1) análisis del sentimiento a nivel global y (2) análisis del sentimiento a nivel de aspectos sobre tuits futbolísticos y de política. El rendimiento obtenido por esta red de aprendizaje profundo es comparado con el de nuestro sistema del año pasado, una regresión logística con una regularización cuadrática. Los resultados experimentales muestran que es necesario incluir estrategias como pre-entrenamiento no supervisado, técnicas específicas para representar palabras como vectores o modificar la arquitectura actual para alcanzar resultados acordes con el estado del arte.

**Palabras clave:** deep learning, long short-term memory, análisis del sentimiento, Twitter

**Abstract:** This paper describes the participation of the LyS group at TASS 2015. In this year's edition, we used a long short-term memory neural network to address the two proposed challenges: (1) sentiment analysis at a global level and (2) aspect-based sentiment analysis on football and political tweets. The performance of this deep learning approach is compared to our last-year model, based on a square-regularized logistic regression. Experimental results show that strategies such as unsupervised pre-training, sentiment-specific word embedding or modifying the current architecture might be needed to achieve state-of-the-art results.

**Keywords:** deep learning, long short-term memory, sentiment analysis, Twitter

## 1 *Introduction*

The 4th edition of the TASS workshop addresses two of the most popular tasks on *sentiment analysis* (SA), focusing on Spanish tweets: (1) polarity classification at a global level and (2) a simplified version of aspect-based sentiment analysis, where the goal is to predict the polarity of a set of predefined and identified aspects (Villena-Román et al., b).

The challenge of polarity classification has been typically tackled from two different angles: lexicon-based and machine learning (ML) approaches. The first group relies on sentiment dictionaries to detect the subjective words or phrases of the text, and defines lexical- (Brooke, Tofiloski, and Taboada, 2009; Thelwall et al., 2010) or syntactic-based rules (Vilares, Alonso, and Gómez-Rodríguez, 2015c) to deal with phenomena such as negation, intensification or irrealis.

The second group focuses on training classifiers through supervised learning algorithms that are fed a number of features (Pang, Lee, and Vaithyanathan, 2002; Mohammad, Kiritchenko, and Zhu, 2013; Hurtado and Pla, 2014). Although competitive when labelled data is provided, they have shown weakness when interpreting the compositionality of complex phrases (e.g. adversative subordinate clauses). In this respect, some studies have evaluated the impact of syntactic-based features on these supervised learning techniques (Vilares, Alonso, and Gómez-Rodríguez, 2015b; Joshi and Penstein-Rosé, 2009) or other related tasks, such as multi-topic detection on tweets (Vilares, Alonso,

[*] This research is supported by the Ministerio de Economía y Competitividad y FEDER (FFI2014-51978-C2) and Xunta de Galicia (R2014/034). The first author is funded by the Ministerio de Educación, Cultura y Deporte (FPU13/01180).

and Gómez-Rodríguez, 2015a).

More recently, deep learning (Bengio, 2009) has shown its competitiveness on polarity classification. Bespalov et al. (2011) introduce a word-embedding approach for higher-order n-grams, using a multi-layer perceptron and a linear function as the output layer. Socher et al. (2013) introduce a new deep learning architecture, a Recursive Neural Tensor Network, which improved the state of the art on the Pang and Lee (2005) movie reviews corpus, when trained together with the Stanford Sentiment Treebank. Tang et al. (2014) suggest that currently existing word embedding methods are not adequate for SA, because words with completely different sentiment might appear in similar contexts (e.g. *'good'* and *'bad'*). They pose an sentiment-specific words embedding (SSWE) model, using a deep learning architecture trained from massive distant-supervised tweets. For Spanish, Montejo-Raéz, García-Cumbreras, and Díaz-Galiano (2014) apply word embedding using Word2Vec (Mikolov et al., 2013), to then use those vectors as features for traditional machine learning techniques.

In this paper we also rely on a deep learning architecture, a long short-term memory (LSTM) recurrent neural network, to solve the challenges of this TASS edition. The results are compared with respect to our model for last year's edition, a logistic regression approach fed with hand-crafted features.

## *2 Task1: Sentiment Analysis at a global level*

Let $L$=$\{l_0, l_1, ..., l_n\}$ be the set of polarity labels and $T$=$\{t_0, t_1, ..., t_m\}$ the set of labelled texts, the aim of the task consists of defining an hypothesis function, $h : T \rightarrow L$.

To train and evaluate the task, the collection from TASS-2014 (Villena-Román et al., 2015) was used. It contains a training set of 7 128 tweets, intended to build and tune the models, and two test sets: (1) a pooling-labelled collection of 60 798 tweets and (2) a manually-labelled test set of 1 000 tweets. The collection is annotated using two different criteria. The first one considers a set of 6 polarities ($L_6$): *no opinion* (NONE), *positive* (P), *strongly positive* (P+), *negative* (N), *strongly negative* (N+) and *mixed* (NEU), that are tweets that mix both negative and positive ideas. A simplified version with 4 classes ($L_4$) is also proposed, where the polarities P+ and N+ are included into P and N, respectively.

In the rest of the paper, we will use $h_4$ and $h_6$ to refer our prediction models for 4 and 6 classes, respectively.

## *3 Task2: Sentiment Analysis at the aspect level*

Let $L$=$\{l_0, l_1, ..., l_n\}$ be the set of polarity labels, $A$=$\{a_0, a_1, ...a_o\}$ the set of aspects and a $T$=$\{t_0, t_1, ..., t_m\}$ the set of texts, the aim of the task consists of defining an hypothesis function, $h : A \times T \rightarrow L$. Two different corpora are provided to evaluate this task: a SOCIAL-TV corpus with football tweets (1 773 training and 1 000 test tweets) and a political corpus (784 training and 500 test tweets), called STOMPOL. Each aspect can be assigned the P, N or NEU polarities ($L_3$).

The TASS organisation provided both $A$ and the identification of the aspects that appear in each tweet, so the task can be seen as identifying the scope $s(a, t)$ of an aspect $a$ in the tweet $t \in T$, with $s$ a substring of $t$ and $a \in A$, to then predict the polarity using the hypothesis function, $h_3(s) \rightarrow L_3$.

To identify the scope we followed a naïve approach: given an aspect $a$ that appears at position $i$ in a text, $t$=$[w_0, ..., w_{i-x}, ..., a_i, ..., w_{i+x}, ..., w_p]$, we created a snippet of length $x$ that is considered to be the scope of the aspect. Preliminary experiments on the SOCIAL-TV and the STOMPOL corpus showed that $x = 4$ and taking the entire tweet were the best options for each collection, respectively.

## *4 Supervised sentiment analysis models*

Our aim this year was to compare our last-year model to a deep learning architecture that was initially available for binary polarity classification.

### 4.1 Long Short-Term Memory

Long Short-Term Memory (LSTM) is a recurrent neural network (RNN) proposed by Hochreiter and Schmidhuber (1997). Traditional RNN were born with the objective of being able to store representations of inputs in form of activations, showing temporal capacities and helping to learn short-term dependencies. However, they might suffer from the problem of *exploding gradients*[1]. The LSTM tries to solve these problems using a different type of units, called *memory cells*, which can remember a value for an arbitrary period of time.

In this work, we use a model composed of a single LSTM and a logistic function as the output layer, which has an available implementation[2] in Theano (Bastien et al., 2012).

To train the model, the tweets were tokenised (Gimpel et al., 2011), lemmatised (Taulé, Martí, and Recasens, 2008), converted to lowercase to reduce sparsity and finally indexed. To train the LSTM-RNN, we relied on ADADELTA (Zeiler, 2012), an adaptive learning rate method, using stochastic training (batch size = 16) to speed up the learning process. Experiments with non-stochastic training runs did not show an improvement in terms of accuracy. We empirically explored the size of the word embedding[3] and the number of words to keep in the vocabulary[4], obtaining the best performance using a choice of 128 and 10 000 respectively.

## 4.2 L2 logistic regression

Our last-year edition model relied on the simple and well-known squared-regularised logistic regression (L2-LG), that performed very competitively for all polarity classification tasks. A detailed description of this model can be found in Vilares et al. (2014a), but here we just list the features that were used: *lemmas* (Taulé, Martí, and Recasens, 2008), *psychometric properties* (Pennebaker, Francis, and Booth, 2001) and *subjective lexicons* (Saralegi and San Vicente, 2013). This architecture also obtained robust and competitive performance for English tweets, on SemEval 2014 (Vilares et al., 2014b).

### Penalising neutral tweets

Previous editions of TASS have shown that the performance on NEU tweets is much lower than for the rest of the classes (Villena-Román et al., a). This year we proposed a small variation on our L2-LG model: a penalising system for NEU tweets to determine the polarities under the $L_6$ configuration, where: given an $L_4$ and an $L_6$ LG-classifier and a tweet $t$, if $h_6(t) =$ NEU and $h_4(t) \neq$ NEU then $h_6(t) := h_4(t)$. The results obtained on the test set shown that we obtained an improvement of 1 percentage point with this strategy (from 55.2% to 56.8% that is reported in the Experiments section).

## 5 *Experimental results*

Table 1 compares our models with the best performing run of the rest of the participants (out of date runs are not included). The performance of our current deep learning model is still far from the top ranking systems, and from our last-year model too, although it worked acceptably under the $L_6$ manually-labelled test.

Table 2 and 3 show the F1 score for each polarity, for the LSTM-RNN and L2-LG models, respectively. The results reflect the lack of capacity of the current LSTM model to learn the minority classes in the training data (P, N+ and NEU). In this respect, we plan to explore how balanced corpora and bigger corpora can help diminish this problem.

| System | Ac 6 | Ac 6-1k | Ac 4 | Ac 4-1k |
|---|---|---|---|---|
| LIF | $0.672_{1}$ | $0.516_{1}$ | $0.725_{1}$ | $0.692_{1}$ |
| ELiRF | $0.659_{2}$ | $0.488_{3}$ | $0.722_{2}$ | $0.645_{5}$ |
| GSI | $0.618_{3}$ | $0.487_{4}$ | $0.690_{4}$ | $0.658_{3}$ |
| DLSI | $0.595_{4}$ | $0.385_{14}$ | $0.655_{6}$ | $0.637_{7}$ |
| GTI-GRAD | $0.592_{5}$ | $0.509_{2}$ | $0.695_{3}$ | $0.674_{2}$ |
| **LYS-LG**$^{\bullet}$ | $\mathbf{0.568}_{6}$ | $\mathbf{0.434}_{5}$ | $\mathbf{0.664}_{5}$ | $\mathbf{0.634}_{9}$ |
| DT | $0.557_{7}$ | $0.408_{10}$ | $0.625_{7}$ | $0.601_{11}$ |
| ITAINNOVA | $0.549_{8}$ | $0.405_{11}$ | $0.610_{10}$ | $0.484_{14}$ |
| BittenPotato | $0.535_{9}$ | $0.418_{8}$ | $0.602_{11}$ | $0.632_{10}$ |
| **LYS-LSTM**$^{\bullet}$ | $\mathbf{0.505}_{9*}$ | $\mathbf{0.430}_{6*}$ | $\mathbf{0.599}_{11*}$ | $\mathbf{0.605}_{10*}$ |
| SINAI-ESMA | $0.502_{10}$ | $0.411_{9}$ | - | - |
| CU | $0.495_{11}$ | $0.419_{7}$ | $0.481_{13}$ | $0.600_{12}$ |
| INGEOTEC | $0.488_{12}$ | $0.431_{6}$ | - | - |
| SINAI | $0.474_{13}$ | $0.389_{13}$ | $0.619_{8}$ | $0.641_{6}$ |
| TID-SPARK | $0.462_{14}$ | $0.400_{12}$ | $0.594_{12}$ | $0.649_{4}$ |
| GAS-UCR | $0.342_{15}$ | $0.338_{15}$ | $0.446_{14}$ | $0.556_{13}$ |
| UCSP | $0.273_{16}$ | - | $0.613_{9}$ | $0.636_{8}$ |

Table 1: Comparison of accuracy for Task 1, between the best performance of each participant with respect to our machine- and deep learning models. Bold runs indicate our L2-LG and LSTM runs. Subscripts indicate the ranking for each group for their best run.

Finally, Table 4 compares the performance of the participating systems Task 2, both for

[1] The gradient signal becomes either too small or large causing a very slow learning or a diverging situation, respectively.

[2] http://deeplearning.net/tutorial/

[3] The size of the vector obtained for each word and the number of hidden units on the LSTM layer.

[4] Number of words to be indexed. The rest of the words are set to *unknown tokens*, giving to all of them the same index.

| Corpus | N+ | N | NEU | NONE | P | P+ |
|---|---|---|---|---|---|---|
| $L_6$ | 0.000 | 0.486 | 0.000 | 0.582 | 0.049 | 0.575 |
| $L_6$-1k | 0.090 | 0.462 | 0.093 | 0.508 | 0.209 | 0.603 |
| $L_4$ | - | 0.623 | 0.00 | 0.437 | 0.688 | - |
| $L_4$-1k | - | 0.587 | 0.00 | 0.515 | 0.679 | - |

Table 2: F1 score of our LSTM-RNN model for each test set proposed at Task 1. *1k* refers to the manually-labelled corpus containing 1 000 tweets.

| Corpus | N+ | N | NEU | NONE | P | P+ |
|---|---|---|---|---|---|---|
| $L_6$ | 0.508 | 0.464 | 0.135 | 0.613 | 0.205 | 0.682 |
| $L_6$-1k | 0.451 | 0.370 | 0.000 | 0.446 | 0.232 | 0.628 |
| $L_4$ | - | 0.674 | 0.071 | 0.569 | 0.747 | - |
| $L_4$-1k | - | 0.642 | 0.028 | 0.518 | 0.714 | - |

Table 3: F1 score of our L2-LG model for each test set proposed at Task 1

football and political tweets. The trend remains in this case and the machine learning approaches outperformed again our deep learning proposal.

## 6 Conclusions and future research

In the 4th edition of TASS 2015, we have tried a long short-term memory neural network to determine the polarity of tweets at the global and aspect levels. The performance of this model has been compared with the performance of our last-year system, based on an L2 logistic regression. Experimental results suggest that we need to explore new architectures and specific word embedding representations to obtain state-of-the-art results on sentiment analysis tasks. In this respect, we believe sentiment-specific word embeddings and other deep learning approaches (Tang et al., 2014) can help enrich our current model. Unsupervised pretraining has also been shown to improve performance of deep learning architectures (Severyn and Moschitti, 2015).

| System | SOCIAL-TV | STOMPOL |
|---|---|---|
| ELiRF | $0.633_1$ | $0.655_1$ |
| LYS-LG• | $\mathbf{0.599}_2$ | $\mathbf{0.610}_4$ |
| GSI | - | $0.635_2$ |
| TID-SPARK | $0.557_3$ | $0.631_3$ |
| LYS-LSTM• | $\mathbf{0.540}_{3*}$ | $\mathbf{0.522}_{4*}$ |

Table 4: Comparison of accuracy for Task 2, between the best run of the rest of participants and our machine and deep learning models

## References

Bastien, F., P. Lamblin, R. Pascanu, J. Bergstra, I. Goodfellow, A. Bergeron, N. Bouchard, D. Warde-Farley, and Y. Bengio. 2012. Theano: new features and speed improvements. *arXiv preprint arXiv:1211.5590.*

Bengio, Y. 2009. Learning deep architectures for AI. *Foundations and trends in Machine Learning*, 2(1):1–127.

Bespalov, D., B. Bai, Y. Qi, and A. Shokoufandeh. 2011. Sentiment classification based on supervised latent n-gram analysis. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 375—-382. ACM.

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

Gimpel, K, N Schneider, B O'connor, D Das, D Mills, J Eisenstein, M Heilman, D Yogatama, J Flanigan, and N A Smith. 2011. 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.

Hochreiter, S and J. Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hurtado, L. and F. Pla. 2014. ELiRF-UPV en TASS 2014: Análisis de sentimientos, detección de tópicos y análisis de sentimientos de aspectos en twitter. In *Proceedings of the TASS workshop at SEPLN.*

Joshi, M 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, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781.*

Mohammad, S. M, S. Kiritchenko, and X. Zhu. 2013. NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*, Atlanta, Georgia, USA, June.

Montejo-Raéz, A., M. A. García-Cumbreras, and M. C. Díaz-Galiano. 2014. Participación de SINAI word2vec en TASS 2014. In *Proceedings of the TASS workshop at SEPLN.*

Pang, B. and L. Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 115–124. Association for Computational Linguistics.

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

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

Saralegi, X. and I. San Vicente. 2013. Elhuyar at TASS 2013. 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 143–150, Madrid, Spain, September.

Severyn, A. and A. Moschitti. 2015. UNITN: Training Deep Convolutional Neural Network for Twitter Sentiment Classification. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 464–469, Denver, Colorado. Association for Computational Linguistics.

Socher, R., A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *EMNLP 2013. 2013 Conference on Empirical Methods in Natural Language Processing. Proceedings of the Conference*, pages 1631–1642, Seattle, Washington, USA. ACL.

Tang, D., F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1555–1565.

Taulé, M., M. A. Martí, and M. Recasens. 2008. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 96–101, Marrakech, Morocco.

Thelwall, M., K. Buckley, G. Paltoglou, D. Cai, and A. Kappas. 2010. Sentiment Strength Detection in Short Informal Text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558, December.

Vilares, D., M. A. Alonso, and C. Gómez-Rodríguez. 2015a. A linguistic approach for determining the topics of Spanish Twitter messages. *Journal of Information Science*, 41(2):127–145.

Vilares, D., M. A. Alonso, and C. Gómez-Rodríguez. 2015b. 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.

Vilares, D., M A Alonso, and C. Gómez-Rodríguez. 2015c. A syntactic approach for opinion mining on spanish reviews. *Natural Language Engineering*, 21(01):139–163.

Vilares, D., Y. Doval, M. A. Alonso, and C. Gómez-Rodríguez. 2014a. LyS at TASS 2014: A prototype for extracting and analysing aspects from spanish tweets. In *Proceedings of the TASS workshop at SEPLN.*

Vilares, D., M. Hermo, M. A. Alonso, C. Gómez-Rodríguez, and Y. Doval.

2014b. LyS : Porting a Twitter Sentiment Analysis Approach from Spanish to English na. In *Proceedings og The 8th InternationalWorkshop on Semantic Evaluation (SemEval 2014)*, number SemEval, pages 411–415.

Villena-Román, J., J. García-Morera, C. Moreno-García, S. Lana-Serrano, and J. C. González-Cristóba. TASS 2013 — a second step in reputation analysis in Spanish. *Procesamiento del Lenguaje Natural*, pages 37–44.

Villena-Román, Julio, Janine García-Morera, Miguel A. García-Cumbreras, Eugenio Martínez-Cámara, M. Teresa Martín-Valdivia, and L. Alfonso Ureña López. Overview of TASS 2015.

Villena-Román, L., E. Martínez-Cámara, Janine Morera-García, and S. M. Jiménez-Zafra. 2015. TASS 2014-the challenge of aspect-based sentiment analysis. *Procesamiento del Lenguaje Natural*, 54:61–68.

Zeiler, M.D. 2012. ADADELTA: An adaptive learning rate method. *arXiv preprint arXiv:1212.5701.*