ELiRF-UPV at SemEval-2018 Tasks 1 and 3: Affect and Irony Detection in Tweets

José-Ángel González, Lluís-F. Hurtado, Ferran Pla
Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València
Camí de Vera, sn
46022, València
{jogonba2, lhurtado, fpla}@dsic.upv.es

Abstract

This paper describes the participation of ELiRF-UPV team at tasks 1 and 3 of Semeval-2018. We present a deep learning based system that assembles Convolutional Neural Networks and Long Short-Term Memory neural networks. This system has been used with slight modifications for the two tasks addressed both for English and Spanish. Finally, the results obtained in the competition are reported and discussed.

1 Introduction

The study of figurative language and affective information expressed in texts is of great interest in sentiment analysis applications because they can change the polarity of a message. The objective of tasks 1 and 3 of Semeval 2018 is the study of these phenomena on Twitter.

Task 1 (Mohammad et al., 2018) is related to Affect in Tweets. Systems have to automatically determine the intensity of emotions and intensity of sentiment or valence of the tweeters from their tweets. The task is divided in five subtasks: emotion intensity regression (EI-Reg), emotion intensity ordinal classification (EI-Oc), sentiment intensity regression (V-Reg), sentiment analysis ordinal classification (V-Oc) and emotion classification (E-C).

Task 3 (Van Hee et al., 2018) addresses the problem of Irony detection in English Tweets. It consists of two subtasks. The first subtask is a two-class (or binary) classification task where the system has to predict whether a tweet is ironic or not. The second subtask is a multiclass classification task where the system has to predict one out of four labels describing i) verbal irony realized through a polarity contrast, ii) verbal irony without such a polarity contrast (i.e., other verbal irony), iii) descriptions of situational irony, iv) non-irony.

This paper describes the main characteristics of the developed system by the ELiRF-UPV team for tasks 1 and 3. We addressed all subtasks of task 1 both for English and Spanish, and all subtasks of task 3.

2 Data Preprocessing

In this work we have taken into account different aspects when preprocessing the tweets. First we removed the accents and converted all the text to lowercase. In general, *emoticons*, *web links*, *hashtags*, *numbers*, and *user mentions*, were substituted by generic tokens. For instance, "#hashtag" $\rightarrow$ "hashtag", ☺ $\rightarrow$ "Slightly Smiling Face", etc. After that, we used TweetMotif (Krieger and Ahn, 2010) as tweet tokenizer, moreover we adapted it to work with Spanish tweets.

3 Resources

On the one hand, for English, we used the following polarity/emotion lexicons: AFFIN (Nielsen, 2011), Bing Liu's Opinion Lexicon (Hu and Liu, 2004), MPQA (Wilson et al., 2005), Sentiment140 (Go et al., 2009), SentiWordnet (Baccianella et al., 2010), NRC Emotion Lexicon (Mohammad and Turney, 2013), NRC Hashtag Emotion Lexicon (Mohammad, 2012) and LIWC2007 (Pennebaker et al., 2014). We also used Word2Vec embeddings (Mikolov et al., 2013a) (Mikolov et al., 2013b) pre-trained by (Godin et al., 2015) with 400 million English tweets.

On the other hand, for Spanish, we used the following polarity/emotion lexicons: ElHPolar (Saralegi and San Vicente, 2013), ISOL (Molina-González et al., 2013), and MLSenticon (Cruz et al., 2014). In addition, we also pre-trained Word2Vec embeddings from 87 million Spanish tweets collected by our team by means of a twitter crawler. In this case, it is a skip-gram architecture

Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), pages 565–569
New Orleans, Louisiana, June 5–6, 2018.

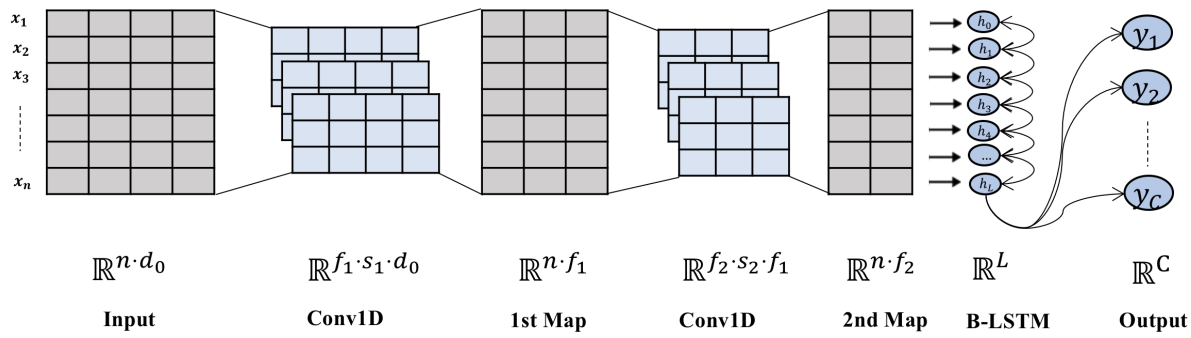

Figure 1: System architecture.

with 300 dimensions per word, negative sampling with 5 negative samples and a 5-term context on the left and right was used.

Through a tuning process on the development sets with a fixed system architecture, we selected the best lexicons for each task. For English, all the lexicons stated above for both tasks were used. For Spanish, only ElHPolar and ISOL lexicons were used.

4 System Description

In this section, we briefly describe the general characteristics of the system developed for Task 1 and Task 3 at SemEval 2018. This description includes the input representation and the system architecture.

4.1 Input representation

Regarding the representation used, in those subtasks where the input is only a tweet (V-Reg, V-Oc and E-C in task 1 and both subtasks in task 3), each tweet is represented as a matrix $M \in \mathbb{R}^{n \cdot d}$ where n is the maximum number of words per tweet and d is the embedding dimensionality. To include the information from the polarity lexicons, for each word, x, the vector of its embedding is concatenated with the vector of polarities/emotions for this word, $v(x)$. In this way, the representation matrix of a tweet finally results in $M \in \mathbb{R}^{n \cdot (d+|v|)}$.

For the EI-Reg subtask, where in addition to a tweet an emotion p is also provided, we add the representation of the word p as the last row of the M matrix. Moreover, we concatenate one column to the word embeddings to indicate if the words belong to the tweet (0) or belong to the emotion (1).

4.2 System architecture

We propose a general architecture for all subtasks. This architecture is based on a two-layer Convolutional Neural Network (CNN) (Fukushima, 1980) ensembled with a final Long Short-Term Memory (LSTM) neural network (Hochreiter and Schmidhuber, 1997) as in (González et al., 2017). We use the representation of the tweet in terms of the M matrix defined above as input to the system. Finally, a fully connected layer computes the outputs of the system. The activation function of this layer depends on the subtask.

Figure 1 shows the general architecture of the system, where d_0 is the dimensionality of the representation of each word (size of the embedding), f_i is the number of filters in the convolutional layer i, s_i the height of each filter in layer i, L is the dimensionality of the *output-state* of the LSTM network, and C is the number of outputs for a specific task.

Although the architecture was the same for all subtasks, the parameters are subtask dependent and were experimentally defined by means of a tuning phase with the development sets. The values studied for the parameters of the convolutional network are $f_i \in [64, 256]$ and $s_i = 3$. The number of neurons of the last layer depends on the subtask. We also tested a simplified version of the architecture without the convolutional network and using only the LSTM network with $L = 256$.

Moreover, we use Batch Normalization (Ioffe and Szegedy, 2015) between all convolutional layers, Dropout (Srivastava et al., 2014) after the LSTM with $p = 0.2$, ReLU activation functions (Nair and Hinton, 2010) and RMSProp (Tieleman and Hinton) as optimization algorithm.

Task 1	EI-Reg (Pearson)		V-Reg (Pearson)		E-C (Jaccard)	
	En	**Sp**	**En**	**Sp**	**En**	**Sp**
LSTM + Lexicons (MSE)	**67.57**	**68.98**	75.46	74.37	N/A	N/A
CNN-LSTM + Lexicons (MSE)	64.12	66.56	**81.13**	**80.01**	N/A	N/A
CNN-LSTM + Lexicons (CCE)	N/A	N/A	N/A	N/A	52.11	42.18
CNN-LSTM + Lexicons (Jaccard)	N/A	N/A	N/A	N/A	**55.23**	**44.59**

Table 1: Task 1 development results.

Task 3	Subtask A (F1)	Subtask B (Macro F1)
CNN-LSTM + Lexicons (CCE)	68.44	44.59
CNN-LSTM + Lexicons (F1)	**68.63**	N/A
CNN-LSTM + Lexicons (Macro F1)	N/A	**45.45**

Table 2: Task 3 development results.

Regarding the loss function, we used *Mean Squared Error* (MSE) for the regression subtasks. However, for subtask E-C and both subtasks of task 3, we used an adaptation of the evaluation metrics (*Jaccard Index*, F_1 for binary classification, and macro-average F_1) as loss functions. In future work we will define and study in more detail this kind of loss functions. In addition, we also tested Cross Entropy (CCE) to extend the comparison.

The strategy used in the ordinal classification subtasks of task 1 (EI-Oc and V-Oc) consisted in the discretization of the outputs of the equivalent regression subtasks (EI-Reg and V-Reg). The discretization process is as follows, be $\mathbb{C}$ the classes set of a ordinal classification subtask and $v_x \in \mathbb{R}$ the score assigned to sample x using a regression model. We compute $|\mathbb{C}| + 1$ thresholds by searching the minimum output for each class, according to the regression train sets. Concretely, $\{th_0, ..., th_{|\mathbb{C}|}\}$ where $th_i \in \mathbb{R}$, $th_i < th_{i+1}$, $th_0 = 0$, and $th_{|\mathbb{C}|} = 1$. Sample x is assigned to the class i such that $th_i < v_x \leq th_{i+1}$.

5 Experimental Results

We performed a tuning process with the development sets in order to select the best model for each task. We tested different ways of preprocessing the tweets, we fit the parameters of the models and we evaluated some external lexicons. Next, we summarize the best results obtained in the tuning process by considering some combinations of the tested models and configurations.

Table 3 shows the results for 3 of the subtasks in the tuning process for Task 1. For the two remaining tasks (EI-Oc and V-Oc) we do not learn specific models, in these cases we used the best models obtained for EI-Reg and V-Reg, respectively.

As it can be seen, LSTM achieved the best results for subtasks El-Reg. The rest of subtasks performed better when we combined CNN and LSTM models. In addition, when we consider the evaluation metric as loss function we improved the results (see the differences between CNN-LSTM + Lexicons (Jaccard) and CNN-LSTM + Lexicons (CCE)).

Table 3 shows the results for the two subtasks in the tuning process for Task 3. We can observe the same behavior as Task 1. The best results are obtained using a combination of CNN and LSTM models and if we consider the evaluation metric as loss function the results are improved.

Once our best system for each subtask with the development set was chosen, we tested it on the official test set and we compare it with the best results obtained by another participant. These results are shown in Table 5 for Task 1, and in Table 5 for Task 3.

Task 1	English		Spanish	
	Our	**Best**	**Our**	**Best**
EI-Reg	69.60(13/42)	79.90	64.80(3/12)	73.80
EI-Oc	59.00(10/36)	69.50	57.50(4/13)	66.40
V-Reg	80.40(15/33)	87.30	74.20(2/12)	79.50
V-Oc	75.90(12/34)	83.60	72.90(2/11)	75.60
E-C	55.20(9/35)	58.80	45.80(2/14)	46.90

Table 3: Task 1 test results.

Task 3	Our	Best
Subtask A	62.94 (7/44)	70.54
Subtask B	42.11 (8/32)	50.74

Table 4: Task 3 test results.

6 Conclusions and Future Work

We presented a deep learning based system that assembles CNN and LSTM neural networks for tasks 1 and 3 of Semeval-2018. This system has been used with slight modifications for the two tasks addressed.

We want to highlight the improvements obtained when the evaluation measures have been adapted as loss functions. In addition, we have also incorporated information extracted from different lexical resources into the models.

As future work, we will continue to study different loss functions and the incorporation of new lexical resources as well as to carry out a detailed study of the obtained results.

7 Acknowledgements

This work has been partially supported by the Spanish MINECO and FEDER founds under projects ASLP-MULAN: Audio, Speech and Language Processing for Multimedia Analytics (TIN2014-54288-C4-3-R); and AMIC: Affective Multimedia Analytics with Inclusive and Natural Communication (TIN2017-85854-C4-2-R). Work of José-Ángel González is also financed by Universitat Politècnica de València under grant PAID-01-17.

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *in Proc. of LREC*.

Fermín L. Cruz, José A. Troyano, Beatriz Pontes, and F. Javier Ortega. 2014. Building layered, multilingual sentiment lexicons at synset and lemma levels. *Expert Systems with Applications*, 41(13):5984 – 5994.

Kunihiko Fukushima. 1980. Neocognn: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. 150.

Fréderic Godin, Baptist Vandersmissen, Wesley De Neve, and Rik Van de Walle. 2015. Multimedia lab @ ACL W-NUT NER sharedtask: named entity recognition for Twitter microposts using distributed word representations. *ACL-IJCNLP*, 2015:146–153.

José-Àngel González, Ferran Pla, and Lluís-F. Hurtado. 2017. Elirf-upv at semeval-2017 task 4: Sentiment analysis using deep learning. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 723–727. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

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.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167.

Michel Krieger and David Ahn. 2010. Tweetmotif: exploratory search and topic summarization for twitter. In *In Proc. of AAAI Conference on Weblogs and Social*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

Saif Mohammad. 2012. #emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, Montréal, Canada. Association for Computational Linguistics.

Saif M. Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 Task 1: Affect in tweets. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, USA.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. 29(3):436–465.

M. Dolores Molina-González, Eugenio Martínez-Cámara, María-Teresa Martín-Valdivia, and José M. Perea-Ortega. 2013. Semantic orientation for polarity classification in spanish reviews. *Expert Systems with Applications*, 40(18):7250 – 7257.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 807–814, USA. Omnipress.

F. Å. Nielsen. 2011. AFINN.

JW Pennebaker, CK Chung, M Ireland, A Gonzales, and RJ Booth. 2014. The development and psychological properties of liwc2007.

Xabier Saralegi and Inaki San Vicente. 2013. Elhuyar at tass 2013. In *XXIX Congreso de la Sociedad Espaola de Procesamiento de lenguaje natural, Workshop on Sentiment Analysis at SEPLN (TASS2013)*, pages 143–150.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

T. Tieleman and G. Hinton. RMSprop Gradient Optimization.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. Semeval-2018 Task 3: Irony detection in English Tweets. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, USA.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354, Stroudsburg, PA, USA. Association for Computational Linguistics.