TASS 2017: Workshop on Semantic Analysis at SEPLN, septiembre 2017, págs. 29-34

ELiRF-UPV en TASS 2017: Análisis de Sentimientos en Twitter basado en Aprendizaje Profundo

ELiRF-UPV at TASS 2017: Sentiment Analysis in Twitter based on Deep Learning

Lluís-F. Hurtado, Ferran Pla, José-Ángel González
Universitat Politècnica de València
Camí de Vera s/n
46022 València
{lhurtado, fpla, jogonba2}@dsic.upv.es

Resumen: En este trabajo se describe la participación del equipo del grupo de investigación ELiRF de la Universitat Politècnica de València en el Taller TASS2017. Este taller es un evento enmarcado dentro de la XXXIII edición del Congreso Internacional de la Sociedad Española para el Procesamiento del Lenguaje Natural. Este trabajo presenta las aproximaciones utilizadas para las todas las tareas del taller, los resultados obtenidos y una discusión de los mismos. Nuestra participación se ha centrado principalmente en explorar diferentes aproximaciones basadas en *deep learning*, consiguiendo resultados competitivos en las tareas abordadas.
Palabras clave: Twitter, Análisis de Sentimientos, Deep Learning.

Abstract: This paper describes the participation of the ELiRF research group of the Universitat Politècnica de València at TASS2017 Workshop. This workshop is a satellite event of the XXXIII edition of the International Conference of the Spanish Society for Natural Language Processing. This work describes the approaches used for all the tasks of the workshop, the results obtained and a discussion of these results. Our participation has focused primarily on exploring different approaches of *deep learning* and we have achieved competitive results in the addressed tasks.
Keywords: Twitter, Sentiment Analysis, Deep Learning.

1 Introducción

El Taller de Análisis de Sentimientos (TASS) ha venido planteando una serie de tareas relacionadas con el análisis de sentimientos en Twitter con el fin de comparar y evaluar las diferentes aproximaciones presentadas por los participantes. Además, desarrolla recursos de libre acceso, básicamente, corpora anotados con polaridad, temática, tendencia política, aspectos, que son de gran utilidad para la comparación de diferentes aproximaciones a las tareas propuestas.

En esta sexta edición del TASS (Martínez-Cámara et al., 2017) se proponen dos tareas: 1) Determinación de la polaridad de los tweets a nivel global y 2) Determinación de la polaridad a nivel de aspecto. Para la primera tarea se utilizan dos corpus distintos. Un nuevo corpus etiquetado para esta edición, denominado InterTASS, y el *General Corpus* utilizado en ediciones anteriores (Villena-Román et al., 2016). Para la segunda tarea se utilizaron los corpus Social_TV, compuesto por tweets publicados durante la final de la Copa del Rey 2014 y STOMPOL, que consta de un conjunto de tweets sobre diferentes aspectos pertenecientes al dominio de la política.

El presente artículo resume la participación del equipo ELiRF-UPV de la Universitat Politècnica de València en todas las tareas planteadas en este taller. Primero se describen las aproximaciones y recursos utilizados en cada tarea. A continuación se presenta la evaluación experimental realizada y los resultados obtenidos. Finalmente se muestran las conclusiones y posibles trabajos futuros.

2 Descripción de los sistemas

Los sistemas presentados en el TASS 2017 cambian el enfoque utilizado en pasadas ediciones en las que nuestro equipo ha participado (Pla y Hurtado, 2013; Hurtado y Pla, 2014; Hurtado, Pla, y Buscaldi, 2015; Hurtado y Pla, 2016), utilizando modelos basados en SVM entrenados con representaciones

ISSN 1613-0073

bag-of-words de los tweets. En (Pla y Hurtado, 2017) se puede consultar un resumen de nuestras participaciones en el TASS hasta 2016.

El preproceso de los tweets utiliza la estrategia descrita en los trabajos anteriores. Ésta consiste básicamente en la adaptación para el castellano del tokenizador de tweets *Tweetmotif* (O'Connor, Krieger, y Ahn, 2010). La tokenización ha consistido en agrupar todas las fechas, signos de puntuación, números, direcciones web, hashtags y menciones de usuario. Además, se ha considerado y evaluado el uso de palabras y caracteres como tokens.

Todas las tareas se han abordado como un problema de clasificación y para esto se han utilizado redes neuronales debido a que han obtenido buenos resultados en tareas similares (González, Pla, y Hurtado, 2017a). El sistema se ha desarrollado en *Python* y utiliza la librería *Keras* (Chollet y otros, 2015).

En este trabajo se han explorado diferentes topologías de redes neuronales así como diferentes tipos de representaciones de los tweets para obtener los mejores resultados posibles sobre un conjunto de validación extraído para cada tarea. Generalmente se ha experimentado con *Multilayer Perceptron* (MLP) y con redes neuronales recurrentes, típicamente *Long Short Term Memory* (LSTM) (Hochreiter y Schmidhuber, 1997) y *stacks* de redes convolucionales (CNN) y LSTM (González, Pla, y Hurtado, 2017b).

Así, en función de cada modelo se han empleado diversos tipos de representaciones como *bag-of-words*, *bag-of-chars*, colapsado de *embeddings* y representaciones secuenciales formadas por secuencias de *embeddings* o de vectores *one-hot* sobre palabras y caracteres.

Como recursos adicionales, se ha experimentado añadiendo información de polaridad con el lexicón NRC (Mohammad y Turney, 2013), sin embargo, la inclusión de información de polaridad haciendo uso de dicho lexicón no ha mejorado los resultados previamente obtenidos en validación. Con respecto a los *embeddings* utilizados, se ha empleado una arquitectura *skip-gram* (Mikolov et al., 2013a) (Mikolov et al., 2013b) (Řehůřek y Sojka, 2010) con 300 dimensiones para cada *token* de la *lookup table* y un contexto formado por 5 componentes en ambos sentidos, entrenada con 87 millones de *tweets* en español recopilados entre el 1/06/17 y el 14/06/17.

Además, también se ha experimentado con *sentiment specific word embeddings* (SSWE) (Tang, 2015). Puesto que es necesario disponer de datos etiquetados para entrenar este tipo de arquitecturas, hemos aplicado un heurístico basado en la presencia de algunos emoticonos específicos para etiquetar de forma automática un subconjunto de los 87 millones de tweets descritos anteriormente. En total se utilizaron 743807 tweets para aprender los modelos SSWE. Se entrenó una red similar a SSWE-Unified (Tang, 2015) en dos pasos: primero optimizando los *embeddings* con información contextual (*skip-gram*) y posteriormente, en función de la polaridad aproximada determinada aplicando el heurístico mencionada anteriormente.

En resumen, en las experimentaciones realizadas, cada *tweet* ha sido representado de distintas maneras en función del modelo con el que se han realizado dichas experimentaciones y se han estimado los hiper parámetros de los modelos (número de capas, neuronas por capa, iteraciones, etc.), así como la mejor representación para cada modelo, mediante *holdout* con particiones de 80 % para entrenamiento y 20 % para validación en aquellos casos en los que la organización del TASS no proporciona un conjunto de validación.

Para cada tarea, con carácter general, se ha elegido las tres combinaciones de representación de tweets y modelo de clasificación que optimizan la métrica oficial de evaluación propuesta por la organización (*macro-*F_1).

3 Tarea 1: Análisis de sentimientos en tweets

Esta tarea consiste en determinar la polaridad de los tweets. Se trata de un problema de análisis de sentimientos a nivel global utilizando cuatro etiquetas de polaridad: **N** que expresa polaridad negativa, **NEU** y **NONE** para polaridades neutras o ausencia de polaridad respectivamente y **P** para la polaridad positiva.

La organización del TASS ha definido tres subtareas considerando tres corpus diferentes. En primer lugar, el corpus InterTASS compuesto por una partición de entrenamiento de 1008 muestras, una de validación de 506 muestras y otra de test formada por 1920 muestras. En segundo lugar, las dos subtareas restantes utilizan para entrenamiento el *General Corpus* compuesto por 7219 muestras; pero, mientras una utiliza el conjunto de test completo (60798 muestras) la otra utiliza un

subconjunto de 1000 muestras reetiquetadas denominado *General Corpus 1K*.

La distribución de *tweets* según su polaridad en el conjunto de entrenamiento del corpus InterTASS se muestra en la Tabla 1.

Polaridad	# tweets	%
N	418	41.46
NEU	133	13.19
NONE	139	13.78
P	318	31.54
TOTAL	1008	100

Tabla 1: Distribución de tweets en el conjunto de entrenamiento de InterTASS según su polaridad.

Como se puede observar en la Tabla 1, el corpus está desbalanceado predominando las clases **N** (41.46 %) y **P** (31.54 %). Para intentar mitigar este problema, se ha empleado, en toda la experimentación, un escalado de la función de perdida (*loss*) de forma que cada clase tiene asociado un factor multiplicativo a aplicar sobre dicha función. Así, asignándoles un factor mayor a las clases minoritarias, al cometer errores con las muestras de dichas clases la *loss* será mucho mayor, forzando así a las redes neuronales para que clasifiquen correctamente dichas muestras (González, Pla, y Hurtado, 2017a).

Como se ha comentado anteriormente, se realizó un proceso de ajuste donde se determinó el mejor conjunto de entrenamiento, la mejor representación de los tweets y los mejores valores para los hiperparámetros de los modelos. De esta manera, para los sistemas basados en MLP se decidió utilizar como entrenamiento para las tres subtareas el conjunto de entrenamiento de InterTASS, y en los casos donde el sistema está basado en CNN y LSTM se utilizó una combinación de los conjuntos de entrenamiento del InterTASS y del *General Corpus*. Esto es debido a que, como se pudo comprobar en la fase de ajuste, los modelos basados en CNN y LSTM se comportaban mejor con un mayor número de muestras en entrenamiento.

Con ello, para las subtarea del corpus InterTASS, se han considerado los siguientes sistemas:

- **run1**: La primera alternativa está basada en un modelo MLP que colapsa los *embeddings*, extraídos mediante un modelo *skip-gram* entrenado con 87 millones de *tweets*, de las palabras de un *tweet* dado mediante la función suma. El MLP está compuesto por dos capas con funciones de activación ReLU y 128 neuronas en las que se aplica dropout con $p = 0{,}4$ (Srivastava et al., 2014) para mejorar la generalización del sistema, además, se emplea Adagrad (Duchi, Hazan, y Singer, 2011) como algoritmo de optimización con el objetivo de minimizar la entropía cruzada entre la distribución original y la estimada.
- **run2**: El segundo sistema se basa en un *stack* de CNN y LSTM entrenado a partir de los mismos *embeddings* del sistema anterior. En este caso, la red está compuesta por una capa convolucional de 128 *kernels* de anchura 4, un LSTM con 64 neuronas y un MLP formado por dos capas con 64 neuronas y funciones de activación ReLU. También se emplea dropout con $p = 0{,}3$, Batch Normalization (BN) (Ioffe y Szegedy, 2015), Adagrad y se minimiza la entropía cruzada.
- **run3**: La última alternativa es similar al primer sistema, pero empleando *embeddings* específicos de polaridad (SSWE), obtenidos tal como se ha comentado en el apartado anterior.

Para la subtarea sobre el *General Corpus* con el test de 60798 muestras, se han empleado tres sistemas diferentes a los de la subtarea con el corpus InterTASS:

- **run1**: La primera alternativa es el run2 de la subtarea anterior (*stack* de CNN y LSTM) entrenado a partir de *embeddings*, extraídos mediante el modelo *skip-gram* ya mencionado.
- **run2**: La segunda alternativa es similar al primer sistema, pero empleando SSWE.
- **run3**: Con la intención de comparar los nuevos modelos y los modelos utilizados anteriormente por nuestro grupo, basados en SVM, el último sistema presentado ha sido el ganador de la edición de 2016 del TASS (Hurtado y Pla, 2016). Cabe destacar que este modelo fue ajustado utilizando como medida la *Accuracy* que fue la medida oficial en todas las ediciones anteriores.

Sobre la subtarea con el test de 1000 muestras, subconjunto de *General Corpus*, se ha extraído la polaridad de cada muestra a partir de las salidas sobre el test de 60798 muestras, por lo que no se ha desarrollado ningún sistema específico para dicha subtarea.

En la Tabla 2 se muestran los valores de *accuracy* y *macro-*F_1 obtenidos para las tres subtareas. Con los sistemas presentados se obtienen mejoras respecto a los resultados presentados en la edición anterior.

	Run	Acc	F1
InterTASS	run1	**60.70**	**49.30**
	run2	43.60	45.00
	run3	59.70	46.60
General Corpus	run1	66.60	54.20
	run2	65.90	**54.90**
	run3	**72.50**	54.80
General Corpus 1K	run1	**63.00**	**51.90**
	run2	59.60	50.40
	run3	58.80	47.70

Tabla 2: Resultados oficiales del equipo *ELiRF-UPV* en las tres subtareas de la Tarea 1 (TASS-2017).

4 Tarea 2: Análisis de Polaridad de Aspectos en Twitter

Esta tarea consiste en asignar la polaridad a los aspectos que aparecen marcados en el corpus. Una de las dificultades de la tarea consiste en definir qué contexto se le asigna a cada aspecto para poder establecer su polaridad. Para un problema similar, detección de la polaridad a nivel de entidad, en la edición del TASS 2013, propusimos una segmentación de los tweets basada en un conjunto de heurísticas (Pla y Hurtado, 2013). Esta aproximación también se utilizó para la tarea de detección de la tendencia política de los usuarios de Twitter (Pla y Hurtado, 2014) y para este caso proporcionó buenos resultados. En este trabajo se emplea la aproximación utilizada en la edición del TASS 2016, más simple que las mencionadas y que consiste en determinar el contexto de cada aspecto a través de una ventana fija definida a la izquierda y derecha de la instancia del aspecto. La longitud de la ventana óptima se ha determinado experimentalmente sobre el conjunto de entrenamiento mediante *holdout*.

La organización del TASS ha planteado dos subtareas. La primera utiliza el corpus *Social_TV* y la segunda el corpus *STOMPOL*.

4.1 Corpus Social_TV

El corpus Social_TV fue proporcionado por la organización y se compone de un conjunto de tweets recolectados durante la final de la Copa del Rey de fútbol de 2014. Está dividido en 1773 tweets de entrenamiento y 1000 tweets de test. El conjunto de entrenamiento está anotado con los aspectos y su correspondiente polaridad, utilizando en este caso sólo tres valores: P, N y NEU. El conjunto de test está anotado con los aspectos y se debe determinar la polaridad de éstos.

4.2 Corpus STOMPOL

El corpus STOMPOL se compone de un conjunto de tweets relacionados con una serie de aspectos políticos, como economía, sanidad, etc. que están enmarcados en la campaña política de las elecciones andaluzas de 2015. Cada aspecto se relaciona con una o varias entidades que se corresponden con uno de los principales partidos políticos en España (PP, PSOE, IU, UPyD, Cs y Podemos). El corpus consta de 1.284 tweets, y ha sido dividido en un conjunto de entrenamiento (784 tweets) y un conjunto de evaluación (500 tweets).

4.3 Aproximación y resultados

Los sistemas utilizados son idénticos a los empleados con el corpus InterTASS de la primera tarea. Tampoco se han utilizado lexicones de polaridad debido a que no mejoraban resultados sobre validación y se hace uso de *embeddings* obtenidos con *skip-gram* y SSWE. Además, antes de abordar el entrenamiento se determinan los segmentos de *tweet* que constituyen el contexto de cada uno de los aspectos presentes.

Con ello, se han tenido en cuenta diferentes tamaños de contexto y se han escogido los mejores para cada sistema de cada experimentación en función de los resultados obtenidos en validación. Posteriormente, cada segmento se *tokeniza* de la misma manera que en la Tarea 1 y se escoge el mejor modelo mediante validación.

Por un lado, para la tarea con el corpus Social_TV, los sistemas enviados son los siguientes:

- **run1**: La primera alternativa es el run1 de la Tarea 1, MLP con colapsado suma de *embeddings* extraídos con *skip-gram*.

En este caso, el mejor contexto en validación fue el formado por dos *tokens* a izquierda y derecha del aspecto.

- **run2**: La segunda alternativa es la misma red que el run2 de la Tarea 1, *stack* de CNN y LSTM con *embeddings* extraídos con *skip-gram*. En este caso, el mejor contexto fue de tres *tokens* a izquierda y derecha del aspecto.
- **run3**: El último sistema es idéntico al run1 de esta tarea pero empleando SSWE. El mejor contexto también está formado por tres *tokens* en ambos sentidos.

Por otro lado, para la tarea con STOMPOL, los sistemas considerados son idénticos a la anterior tarea, pero con tamaños de contexto diferentes. En concreto, para el run1 y el run2 el tamaño de contexto es de 5 *tokens* a izquierda y derecha y en el run3 de 4 *tokens*.

Por último, los resultados obtenidos de *Accuracy* y *macro-*F_1 con los distintos sistemas para cada subtarea se muestran en la Tabla 3.

	Run	Acc	F1
Social_TV	run1	**62.50**	47.60
	run2	60.00	51.30
	run3	61.50	**53.70**
STOMPOL	run1	**61.50**	**53.70**
	run2	54.10	48.60
	run3	57.80	48.60

Tabla 3: Resultados oficiales del equipo *ELiRF-UPV* en las dos subtareas de la tarea 2 (TASS-2017).

5 Conclusiones y trabajos futuros

En este trabajo se ha presentado la participación del equipo ELiRF-UPV en las 2 tareas planteadas en TASS 2017. Nuestro equipo ha utilizado técnicas de aprendizaje automático, en concreto, aproximaciones basadas en redes neuronales y, en algunos casos *deep learning*. Para ello hemos utilizado la librería *Keras* para *Python* y otras como *Gensim* (Řehůřek y Sojka, 2010) para la obtención de *embeddings* mediante modelos *skip-gram*. Nuestra participación se ha centrado principalmente en explorar diferentes modelos y representaciones de los tweets, consiguiendo mejorar, en algunas tareas, las prestaciones de ediciones anteriores.

Nuestro grupo está interesado en seguir trabajando en la minería de textos en redes sociales (*author profiling*, *stance detection*, *sentiment analysis*, etc.), así como en tareas relacionadas (seguimiento de tendencias: políticas y radicalización, entre otras) y especialmente en incorporar nuevos recursos a los sistemas desarrollados y estudiar nuevas estrategias y métodos de *deep learning*, a destacar los métodos generativos como *Generative Adversarial Networks* (Goodfellow et al., 2014) y *Variational Autoencoders* (Kingma y Welling, 2013) en combinación con modelos *sequence-to-sequence* (Sutskever, Vinyals, y Le, 2014) para tratar el problema del desbalanceo en corpus similares a los tratados.

Agradecimientos

Este trabajo ha sido parcialmente subvencionado por MINECO y fondos FEDER bajo el proyecto ASLP-MULAN: Audio, Speech and Language Processing for Multimedia Analytics, TIN2014-54288-C4-3-R.

Bibliografía

Chollet, F. et al. 2015. Keras. `https://github.com/fchollet/keras`.

Duchi, J., E. Hazan, y Y. Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12:2121–2159, Julio.

González, J.-A., F. Pla, y L.-F. Hurtado. 2017a. ELiRF-UPV at IberEval 2017: Stance and Gender Detection in Tweets. En *Notebook Papers of 2nd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IBEREVAL)*, Murcia, Spain. CEUR Workshop Proceedings. CEUR-WS.org.

González, J.-A., F. Pla, y L.-F. Hurtado. 2017b. ELiRF-UPV at SemEval-2017 Task 4: Sentiment Analysis using Deep Learning. En *Proceedings of the 11th International Workshop on Semantic Evaluation*, SemEval '17, páginas 722–726, Vancouver, Canada, August. Association for Computational Linguistics.

Goodfellow, I. J., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, y Y. Bengio. 2014. Generative adversarial networks. *CoRR*, abs/1406.2661.

Hochreiter, S. y J. Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Hurtado, L.-F. y F. Pla. 2014. Elirf-upv en tass 2014: Análisis de sentimientos, detección de tópicos y análisis de sentimientos de aspectos en twitter. *Procesamiento del Lenguaje Natural*.

Hurtado, L.-F. y F. Pla. 2016. Elirf-upv en tass 2016: Análisis de sentimientos en twitter. En *TASS@SEPLN*.

Hurtado, L.-F., F. Pla, y D. Buscaldi. 2015. Elirf-upv en tass 2015: Análisis de sentimientos en twitter. En *TASS@ SEPLN*, páginas 75–79.

Ioffe, S. y C. Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167.

Kingma, D. P. y M. Welling. 2013. Auto-encoding variational bayes. *CoRR*, abs/1312.6114.

Martínez-Cámara, E., M. C. Díaz-Galiano, M. A. García-Cumbreras, M. García-Vega, y J. Villena-Román. 2017. Overview of TASS 2017. En J. Villena Román M. A. García Cumbreras E. Martínez-Cámara M. C. Díaz Galiano, y M. García Vega, editores, *Proceedings of TASS 2017: Workshop on Semantic Analysis at SEPLN (TASS 2017)*, volumen 1896 de *CEUR Workshop Proceedings*, Murcia, Spain, September. CEUR-WS.

Mikolov, T., K. Chen, G. Corrado, y J. Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Mikolov, T., I. Sutskever, K. Chen, G. Corrado, y J. Dean. 2013b. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

Mohammad, S. M. y P. D. Turney. 2013. Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence*, 29(3):436–465.

O'Connor, B., M. Krieger, y D. Ahn. 2010. Tweetmotif: Exploratory search and topic summarization for twitter. En *ICWSM*, páginas 384–385.

Pla, F. y L.-F. Hurtado. 2013. Elirf-upv en tass-2013: Análisis de sentimientos en twitter. En *XXIX Congreso de la Sociedad Espanola para el Procesamiento del Lenguaje Natural (SEPLN 2013). TASS*, páginas 220–227.

Pla, F. y L.-F. Hurtado. 2014. Political tendency identification in twitter using sentiment analysis techniques. En *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, páginas 183–192, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Pla, F. y L.-F. Hurtado. 2017. Spanish sentiment analysis in twitter at the tass workshop. *Language Resources and Evaluation*, Jun.

Řehůřek, R. y P. Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. En *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, páginas 45–50, Valletta, Malta, Mayo. ELRA. `http://is.muni.cz/publication/884893/en`.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, y R. Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, Enero.

Sutskever, I., O. Vinyals, y Q. V. Le. 2014. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215.

Tang, D. 2015. Sentiment-specific representation learning for document-level sentiment analysis. En *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, páginas 447–452, New York, NY, USA. ACM.

Villena-Román, J., M. Á. G. Cumbreras, E. M. Cámara, M. C. Díaz-Galiano, M. T. Martín-Valdivia, y L. A. U. López, editores. 2016. *Proceedings of TASS 2016: Workshop on Sentiment Analysis at SEPLN co-located with 32nd SEPLN Conference (SEPLN 2016), Salamanca, Spain, September 13th, 2016*, volumen 1702 de *CEUR Workshop Proceedings*. CEUR-WS.org.