TASS 2018: Workshop on Semantic Analysis at SEPLN, septiembre 2018, págs. 103-109

ELiRF-UPV en TASS 2018: Categorización Emocional de Noticias

ELiRF-UPV at TASS 2018: Emotional Categorization of News Articles

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

Resumen: En este trabajo se describe la participación del grupo de investigación ELiRF de la Universitat Politècnica de València en el Taller TASS2018, enmarcado dentro de la XXXIV edición del Congreso Internacional de la Sociedad Española para el Procesamiento del Lenguaje Natural. Presentamos las aproximaciones utilizadas para la tarea "Good Or Bad News? Emotional categorization of news articles" del TASS, 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 la tarea abordada.
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 TASS2018 Workshop which is a satellite event of the XXXIV edition of the International Conference of the Spanish Society for Natural Language Processing. We describe the approaches used for "Good Or Bad News? Emotional categorization of news articles" task, 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 task.
Keywords: Twitter, Sentiment Analysis, Deep Learning.

1 Introducción

El Taller de Análisis Semántico de la SEPLN (TASS) ha propuesto históricamente tareas relacionadas con el análisis de sentimientos, con el objetivo de evaluar los diversos sistemas planteados por los participantes. Para estas tareas, desarrollan recursos lingüísticos de libre acceso como corpora anotados con polaridad, bien a nivel de aspectos o a nivel global.

En esta séptima edición del TASS (Martínez-Cámara et al., 2018), además de las tareas 1 y 2, dedicadas a análisis de sentimiento, la organización plantea las tareas 3 y 4, dedicadas al descubrimiento de conocimiento en documentos médicos y a la categorización emocional de noticias respectivamente.

Con respecto a la tarea 4, la cual es abordada en este artículo, el objetivo consiste en clasificar titulares de noticias como **SAFE** o **UNSAFE**, en función de si se pueden posicionar anuncios en la noticia o no. El criterio de decisión del posicionamiento de noticias puede llegar a ser muy complejo y, en este caso, la organización de la tarea opta por un criterio basado en las emociones y la temática expresadas en la noticia. Concretamente, se considera que una noticia es **SAFE** en caso de que no exprese ninguna emoción negativa ni trate ningún tema controvertido, en cualquier otro caso, la noticia no es segura para posicionar anuncios y se considera **UNSAFE**.

El presente artículo resume la participación del equipo ELiRF-UPV de la Universitat Politècnica de València en todas las subtareas de la tarea 4, que tratan la categorización emocional de noticias con el objetivo de determinar si es seguro posicionar anuncios en ellas.

El resto del artículo se estructura como si-

ISSN 1613-0073

gue: primero se describen el corpus, las aproximaciones y los recursos utilizados en la tarea. A continuación, se presenta la evaluación experimental realizada y los resultados obtenidos. Finalmente, se muestran las conclusiones y posibles trabajos futuros.

2 Corpus

Para la primera edición de la tarea 4, la organización ha construido el corpus SANSE, compuesto por titulares en variantes del español utilizado en España y en diversos países de América Latina. Así, se han considerado diversas variantes del español con el objetivo de que los sistemas sean capaces de afrontar dificultades relacionadas con la diversidad léxica, sintáctica y temática.

El corpus está compuesto por 2000 titulares de noticias, a partir de los cuales, los organizadores han extraído particiones de entrenamiento, validación y test en función de cada subtarea. Para la subtarea 1, partiendo del conjunto de datos completo, que incluye todas las variedades lingüísticas, se proporciona un conjunto de entrenamiento de 1250 muestras, uno de validación de 250 y otro de test de 500 (L1). Además, proponen un conjunto adicional de test compuesto por 13152 titulares (L2). Por otro lado, para la subtarea 2, los conjuntos de entrenamiento y validación están formados únicamente por los titulares en español de España, con 207 y 48 muestras respectivamente. En este caso, el conjunto de test está compuesto por 407 titulares escritos en el resto de variedades lingüísticas.

En las Tablas 1 y 2 se muestra la distribución de las clases **SAFE** y **UNSAFE** para las subtareas 1 y 2 respectivamente. Como se puede observar, en todas las particiones existe un sesgo hacia la clase **UNSAFE**, aunque la magnitud de dicho sesgo difiere entre particiones e.g. en L2 hay el doble de muestras **UNSAFE** que **SAFE**, mientras que en L1 hay un 19.6 % más de muestras **UNSAFE**. En el caso de las particiones de entrenamiento y validación de la subtarea 1, la magnitud del desbalanceo es idéntica i.e. $\frac{490}{1250} = \frac{98}{250}$ y $\frac{760}{1250} = \frac{152}{250}$. En la subtarea 2, todas las particiones tienen un sesgo similar también hacia la clase **UNSAFE**.

	Train	Dev	L1	L2
SAFE	490	98	201	4461
UNSAFE	760	152	299	8692
Σ	1250	250	500	13152

Tabla 1: Distribución de las muestras en la subtarea 1 para las clases **SAFE** y **UNSAFE**.

	Train	Dev	Test
SAFE	80	19	156
UNSAFE	127	29	251
Σ	207	48	407

Tabla 2: Distribución de las muestras en la subtarea 2 para las clases **SAFE** y **UNSAFE**.

3 Descripción de los sistemas

Como baselines se han implementado dos sistemas basados en Support Vector Machine (Cortes y Vapnik, 1995) que hacen uso de representaciones *bag-of-words* de los titulares, a nivel de palabras (BOW) y de caracteres (BOC). Con ello, para mejorar los resultados de los baselines, se han explorado varias arquitecturas *Deep Learning* y representaciones.

La tokenización utilizada consiste en la adaptación para el castellano del tokenizador *Tweetmotif* (O'Connor, Krieger, y Ahn, 2010). Tras la tokenización, se ha llevado a cabo una etapa de preproceso sobre los titulares que consiste en eliminar acentos y convertir a minúsculas.

Con los baselines y el preproceso ya determinado, se han explorado diversas arquitecturas *Deep Learning*, de la misma forma que en nuestra participación en las tareas 1 y 2 del taller. Concretamente, estudiamos Convolutional Neural Network (CNN) (Kim, 2014), Attention Bidirectional Long Short Term Memory (Att-BLSTM) (Zhou et al., 2016) y Deep Averaging Networks (DAN) (Iyyer et al., 2015). Con respecto a las representaciones, se han empleado tipos distintos en función del modelo utilizado en la experimentación. A destacar: BOW, BOC, *word embeddings* de Twitter (TWE) (Hurtado, Pla, y González., 2017) y el modelo de (Cardellino, 2016) (CWE), así como lexicones de polaridad y emociones (LE) (Mohammad y Turney, 2013), (Saralegi y Vicente, 2013), (L. Cruz et al., 2014), (Molina-González et al., 2013). Para llevar a cabo la experimentación con diversos sistemas y representaciones, hemos utilizado las librerías *Keras* (Chollet, 2015), *Scikit-Learn* (Buitinck et al., 2013) y *Gensim* (Řehůřek y Sojka, 2010).

(1) c = **SAFE,** y = **UNSAFE,** $p(y|x) = 99{,}58\,\%$**:** Venezuela: encontraron dos aeronaves que presuntamente estaban vinculadas al narcotráfico

(2) c = **UNSAFE,** y = **SAFE,** $p(y|x) = 97{,}07\,\%$: Doble premio a la memoria histórica en la Berlinale

(3) c = **SAFE,** y = **UNSAFE,** $p(y|x) = 50{,}80\,\%$: ¿Se viene la lluvia?

(4) c = **UNSAFE,** y = **SAFE,** $p(y|x) = 50{,}80\,\%$: Con su "sinceridad" a prueba: China insta a EE.UU. y Corea del Norte a un diálogo urgente

Tabla 3: Ejemplos de dos errores con $p(y|x)$ máximas y mínimas sobre el conjunto de test L1 de la subtarea 1.

(4) c = **SAFE,** y = **UNSAFE,** $p(y|x) = 99{,}58\,\%$**:** La Comisión de Política Exterior de la AN declaró "fin de la integraciónçon Colombia y Brasil

(5) c = **UNSAFE,** y = **SAFE,** $p(y|x) = 99{,}75\,\%$: Petunia, la segunda opción que descartaron la China Suárez y Benjamín Vicuña como nombre de su hija Magnolia

(6) c = **SAFE,** y = **UNSAFE,** $p(y|x) = 50{,}03\,\%$: En alerta por los hinchas rusos del Lokomotiv que hoy llegan a Madrid

(7) c = **UNSAFE,** y = **SAFE,** $p(y|x) = 50{,}02\,\%$: Kim Cattrall, lapidaria: "Sarah Jessica Parker, no necesito tu apoyo en este trágico momento"

Tabla 4: Ejemplos de dos errores con $p(y|x)$ máximas y mínimas sobre el conjunto de test L2 de la subtarea 1.

Para entrenar los modelos basados en redes neuronales, con el objetivo de evitar el impacto del desbalanceo entre las clases **SAFE** y **UNSAFE**, se ha empleado como función de loss una versión ponderada de la entropía cruzada. Concretamente, $L(x) \cdot log(\mu \cdot \frac{n_r}{n_c})$, donde n_r es el número de muestras en la clase mayoritaria (**UNSAFE**) y n_c es el número de muestras en la clase de la muestra x.

Por último, con respecto al criterio de elección del mejor modelo, se ha escogido la arquitectura *Deep Learning* y la representación que mejor se comporta en la partición de validación de la subtarea 1. Una vez determinada la representación y la arquitectura junto con sus hiperparámetros, este mismo sistema se emplea en la subtarea 2.

4 Fase de ajuste

Para estudiar el comportamiento de los diferentes modelos, se realizó un proceso de ajuste. Así, experimentamos con varios sistemas y representaciones sobre la subtarea 1 para escoger el mejor y reutilizarlo en la subtarea 2. En la Tabla 6 se muestran los resultados obtenidos por cada sistema en el conjunto de validación de la subtarea 1. En dicha tabla, S hace referencia al sistema y R al tipo de representación empleada por dicho sistema.

Como se puede observar, SVM con BOW (1-2gramas) obtiene un 73.20 % *Acc* y 69.88 Macro-F_1, mejorando al baseline SVM con BOC (1-10gramas), aunque las diferencias no son significativas a nivel de *Acc*. Además, si en lugar de utilizar representaciones *bag-of-words* de los titulares, empleamos la suma de embeddings TWE con el mismo sistema SVM, conseguimos mejorar en 3.60 % de *Acc* y 5.80 puntos de Macro-F_1, lo que indica que representaciones que capturan contenido semántico de las palabras aportan información importante para la tarea. Esto puede ser debido a la capacidad de los embeddings de agrupar palabras con temática similar, lo que parece ser relevante al determinar si un titular de noticia aborda un tema controvertido o no.

Por otro lado, si en lugar de utilizar SVM como modelo de clasificación, empleamos modelos basados en *Deep Learning* como DAN, CNN o Att-BLSTM conseguimos mejoras de entre 3.6 % y 8.8 % tanto de Macro-F_1 como de *Acc*. Es necesario destacar que las diferencias de *Acc* entre los modelos *Deep Learning* y los baselines son significativas, a pesar de que los intervalos de confianza son muy amplios debido al reducido número de muestras en el conjunto de validación.

Otros aspectos relevantes de la experimentación consisten en la incorporación de lexicones al mejor modelo (DAN + TWE +

(8) c = **SAFE,** y = **UNSAFE,** $p(y|x)$ = 88,82 %**:** Siete millones de bolívares pagan los pacientes renales por una bolsas de sangre

(9) c = **UNSAFE,** y = **SAFE,** $p(y|x)$ = 58,73 %: Bolívar, Cesar, Sierra Nevada y Córdoba, escenarios de reclutamiento de menores en el Caribe

(10) c = **SAFE,** y = **UNSAFE,** $p(y|x)$ = 50,14 %: Vicepresidente García Linera entrega obra deportiva en Bolivia

(11) c = **UNSAFE,** y = **SAFE,** $p(y|x)$ = 50,22 %: Poder Judicial ordenó congelar cuentas bancarias de Alejandro Toledo

Tabla 5: Ejemplos de dos errores con $p(y|x)$ máximas y mínimas sobre el conjunto de test de la subtarea 2.

	S	R	Macro-P	Macro-R	Macro-F_1	Acc
Subtarea 1	SVM	BOW	72.97	69.26	69.88	73.20±5.48
	SVM	BOC	68.02	66.66	67.01	69.60±5.69
	SVM	TWE	75.67	75.85	75.75	76.80±5.22
	DAN (run1)	TWE	85.71	83.81	**84.52**	**85.60**±4.34
	DAN	CWE	82.94	78.34	79.50	81.60±4.79
	DAN	TWE+LE	83.85	83.18	83.48	84.40±4.48
	Att-BLSTM	TWE	79.51	79.17	79.33	80.40±4.91
	CNN	TWE	80.54	78.59	79.27	80.80±4.87
Subtarea 2	DAN (run1)	TWE	84.72	77.22	**78.56**	**81.25**±11.04

Tabla 6: Resultados de los diversos sistemas sobre los conjuntos de validación.

LE) y en la utilización de embeddings preentrenados con datos más similares a los de la tarea (Cardellino, 2016) (DAN + CWE). En el primer caso, los lexicones de polaridad/emociones no parecen aportar información relevante al clasificador. Del mismo modo, la utilización de CWE conlleva reducciones de 5 puntos de Macro F_1 y de 4 % Acc con respecto a los embeddings TWE, lo que resulta contraintuitivo debido a la mayor similitud entre el dominio de la tarea y el de los embeddings CWE.

De todos los sistemas explorados para la subtarea 1, escogemos aquel que maximiza las dos métricas de evaluación, Acc y Macro-F_1 i.e. DAN + TWE. Una vez escogido el mejor sistema, se utiliza en la subtarea 2 entrenando con el conjunto de entrenamiento de dicha subtarea. Con todo ello, generamos dos runs para cada subtarea, un primer run (**run1**) entrenado únicamente con la partición de entrenamiento y un segundo (**run2**) reentrenando el modelo **run1** con las particiones de entrenamiento y validación durante 3 iteraciones más.

La Figura 1 muestra el sistema propuesto en este trabajo, donde x_i representa el embedding de la palabra i, N representa el uso de Batch Normalization (Ioffe y Szegedy, 2015), F la no linealidad, en este caso ReLU, D se refiere al uso de Dropout (Srivastava et al., 2014) con $p = 0{,}3$, $W_1 \in \mathbb{R}^{512*d_e}$ son los pesos de la única capa oculta y d_e la dimensionalidad de los embeddings. Como algoritmo de optimización se ha empleado Adagrad (Duchi, Hazan, y Singer, 2011) con el objetivo de optimizar la versión ponderada de la entropía cruzada.

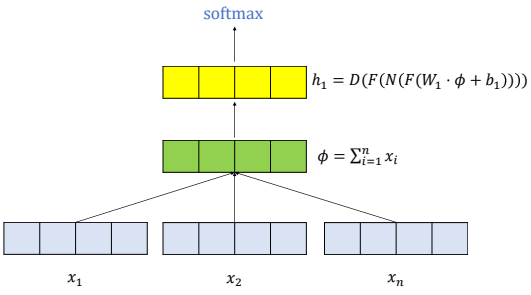

Figura 1: Deep Averaging Network que mejor se comporta en la partición de validación de la subtarea 1.

5 *Resultados experimentales*

En la Tabla 7 se muestran los resultados obtenidos por nuestros sistemas en cada una de las subtareas utilizando los conjuntos de test.

Con ello, en el test L1 de la subtarea 1, nuestros dos sistemas obtienen resultados si-

			Macro-P	Macro-R	Macro-F_1	Acc
Subtarea 1	L1	run1	79.50	78.40	79.00	80.00
		run2	78.70	79.40	79.00	79.40
	L2	run1	87.80	88.90	88.30	89.30
		run2	85.00	88.40	86.70	86.50
Subtarea 2		run1	73.60	64.90	69.00	71.50
		run2	74.70	65.70	69.90	72.20

Tabla 7: Resultados oficiales del equipo *ELiRF-UPV* en la competición (los resultados de los demás participantes se muestran en (Martínez-Cámara et al., 2018)).

milares y observamos cómo considerar la partición de validación durante el entrenamiento nos permite incrementar la Macro-R a costa de reducir Macro-P. Ambos sistemas obtienen el mismo valor de Macro-F_1 pero el **run1** consigue un valor de *Acc* un 0.5 % superior al **run2**. Así, con el sistema **run1** obtenemos un segundo puesto sobre el test L1, 0.5 puntos de macro-F_1 0.2 % de *Acc* por debajo del sistema de INGEOTEC.

Con la partición de test L2 de la subtarea 1, observamos que considerar el conjunto de validación en el entrenamiento nos lleva a una reducción sistemática, en todas las métricas, de hasta un 3 % en el caso de *Acc*. Sobre esta partición, nuestro sistema (**run1**) obtiene los mejores resultados de la competición, tanto a nivel de Macro-F_1 como de *Acc*.

En la subtarea 2, el comportamiento de los sistemas propuestos cambia con respecto a la subtarea 1. Aquí podemos observar como el reentrenamiento con validación (**run2**) permite incrementar los resultados, sobre todas las métricas de evaluación, en aproximadamente 1 punto con respecto al **run1**.

En las Tablas 8, 9 y 10 se muestra la evaluación por clase (precisión, recall y F_1) de los mejores sistemas para cada uno de los conjuntos de test. En ellas se puede observar como, en todos los casos, la clase mejor clasificada es **UNSAFE**, posiblemente debido a la mayor presencia de muestras de esta clase en el corpus. Esto ocurre sobre todas las métricas, excepto en el caso de la precisión en la subtarea 2. En este caso, la precisión del sistema sobre la clase **SAFE** es más alta que sobre **UNSAFE**, sin embargo, esto es a costa de una gran reducción sobre el recall de dicha clase, por lo que el sistema identifica pocas muestras **SAFE** aunque clasifica correctamente la mayoría.

Otro análisis interesante consiste en estudiar las muestras que nuestros mejores sistemas clasifican erróneamente con una gran confianza en la predicción i.e. $\max_y p(y|x)$: $y \neq c(x)$ donde $c(x)$ es la clase correcta para la muestra x e y es la predicción del sistema. También resulta interesante estudiar errores con $p(y|x)$ mínima. Estos ejemplos, sobre las dos tareas y sus respectivos conjuntos de test, se muestran en las Tablas 3, 4, 5.

	P	R	F_1
SAFE	77.90	70.10	73.80
UNSAFE	81.20	86.60	83.80

Tabla 8: Resultados de Precisión, Recall y F_1 por clase para el sistema **run1** en la partición de test L1 de la subtarea 1.

	P	R	F_1
SAFE	82.20	87.50	84.70
UNSAFE	93.40	90.20	91.80

Tabla 9: Resultados de Precisión, Recall y F_1 por clase para el sistema **run1** en la partición de test L2 de la subtarea 1.

En general, observamos en dichas tablas que los valores máximos y mínimos de $p(y|x)$ están cercanos a los valores límite (50 % y 100 %) en todos los conjuntos de test. Sin embargo, en la subtarea 2, los valores máximos son menores en comparación a los test L1 y L2 de la subtarea 1, donde destaca que el segundo valor máximo de $p(y|x)$ (error 9) está más cercano al 50 % que al 100 %. En este caso, se comete el error al predecir la clase **SAFE**, pero sería posible evitarlo si el modelo pudiera observar el trigrama "reclutamiento de menores".

Con respecto a los valores mínimos de $p(y|x)$, en todos los casos están muy próximos al 50 % y algunos de estos errores se pueden abordar mediante la utilización de *Name Entities* (error 4) o reforzando la presencia o ausencia de palabras que emiten emociones negativas (error 7, "lapidaria", "trágico" y error

	P	R	F_1
SAFE	78.70	37.80	51.10
UNSAFE	70.80	93.60	80.60

Tabla 10: Resultados de Precisión, Recall y F_1 por clase para el sistema **run2** en la partición de test de la subtarea 2.

3). Por último, entre todos los errores mostrados en las tablas, observamos algunos que son complejos de clasificar incluso mediante supervisión humana, concretamente, los errores 1 (clase **SAFE**), 2 (clase **UNSAFE**), 4 (clase **SAFE**) y 6 (clase **SAFE**).

6 Conclusiones y trabajos futuros

En este trabajo se ha presentado la participación del equipo ELiRF-UPV en la tarea "Good Or Bad News? Emotional categorization of news articles"planteada en TASS2018. Nuestro equipo ha utilizado modelos *Deep Learning*, obteniendo resultados competitivos en las dos subtareas. Entre ellos, obtenemos los mejores resultados sobre el conjunto de test L2 de la subtarea 1 y un segundo puesto tanto en el conjunto de test L1 de la subtarea 1 como en la subtarea 2.

Durante el desarrollo de la experimentación se han explorado diversas arquitecturas *Deep Learning* y representaciones, con ello, se ha observado que las representaciones basadas en *word embeddings* junto con Deep Averaging Networks aportan mejoras significativas a representaciones y modelos más simples como *bag-of-words* y SVM.

Como trabajo futuro, estamos interesados en mejorar el sistema siguiendo las propuestas planteadas tras el análisis de errores e.g. considerando *Name Entities* o reforzando la presencia o ausencia de palabras que emiten emociones negativas. Además, también resultan de interés otras tareas de minería de textos sobre artículos periodísticos como la detección de *stance*.

Agradecimientos

Este trabajo ha sido parcialmente subvencionado por MINECO y fondos FEDER bajo los proyectos ASLP-MULAN (TIN2014-54288-C4-3-R) y AMIC (TIN2017-85854-C4-2-R). El trabajo de José-Ángel González es también financiado por la Universidad Politécnica de Valencia bajo la beca PAID-01-17.

Bibliografía

Buitinck, L., G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, y G. Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. En *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, páginas 108–122.

Cardellino, C. 2016. Spanish billion words corpus and embeddings. mar.

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

Cortes, C. y V. Vapnik. 1995. Support-vector networks. *Mach. Learn.*, 20(3):273–297, Septiembre.

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.

Hurtado, L.-F., F. Pla, y J.-A. González. 2017. Elirf-upv en TASS 2017: Análisis de sentimientos en twitter basado en aprendizaje profundo. En J. Villena Román M. A. García Cumbreras E. Martínez-Cámara M. C. Díaz Galiano, y M. García Vega, editores, *In Proceedings of TASS 2017: Workshop on Sentiment Analysis at SEPLN co-located with 33nd SEPLN Conference (SEPLN 2017)*, volumen 1896 de *CEUR Workshop Proceedings*, Murcia, Spain, September. CEUR-WS.

Ioffe, S. y C. Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. En *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, páginas 448–456. JMLR.org.

Iyyer, M., V. Manjunatha, J. Boyd-Graber, y H. Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. En *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, páginas 1681–1691. Association for Computational Linguistics.

Kim, Y. 2014. Convolutional neural networks for sentence classification. En *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, páginas 1746–1751. Association for Computational Linguistics.

L. Cruz, F., J. A. Troyano, B. Pontes, y F. J. Ortega. 2014. Building layered, multilingual sentiment lexicons at synset and lemma levels. 41:5984–5994, 10.

Martínez-Cámara, E., Y. Almeida Cruz, M. C. Díaz-Galiano, S. Estévez Velarde, M. A. García-Cumbreras, M. García-Vega, Y. Gutiérrez Vázquez, A. Montejo Ráez, A. Montoyo Guijarro, R. Muñoz Guillena, A. Piad Morffis, y J. Villena-Román. 2018. Overview of TASS 2018: Opinions, health and emotions. En E. Martínez-Cámara Y. Almeida Cruz M. C. Díaz-Galiano S. Estévez Velarde M. A. García-Cumbreras M. García-Vega Y. Gutiérrez Vázquez A. Montejo Ráez A. Montoyo Guijarro R. Muñoz Guillena A. Piad Morffis, y J. Villena-Román, editores, *Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN (TASS 2018)*, volumen 2172 de *CEUR Workshop Proceedings*, Sevilla, Spain, September. CEUR-WS.

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

Molina-González, M. D., E. Martínez-Cámara, M.-T. Martín-Valdivia, y J. M. Perea-Ortega. 2013. Semantic orientation for polarity classification in spanish reviews. *Expert Systems with Applications*, 40(18):7250 – 7257.

O'Connor, B., M. Krieger, y D. Ahn. 2010. Tweetmotif: Exploratory search and topic summarization for twitter.

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`.

Saralegi, X. y I. S. Vicente. 2013. Elhuyar at tweet-norm 2013. En *Proceedings of the Tweet Normalization Workshop colocated with 29th Conference of the Spanish Society for Natural Language Processing (SEPLN 2013), Madrid, Spain, September 20th, 2013.*, páginas 64–68.

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.

Zhou, P., W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, y B. Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. En *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, páginas 207–212. Association for Computational Linguistics.