Participación de SINAI Word2Vec en TASS 2014*

SINAI Word2Vec participation in TASS 2014

A. Montejo-Ráez
University of Jaén
23071 Jaén (Spain)
amontejo@ujaen.es

M.A. García-Cumbreras
University of Jaén
23071 Jaén (Spain)
magc@ujaen.es

M.C. Díaz-Galiano
University of Jaén
23071 Jaén (Spain)
mcdiaz@ujaen.es

Resumen: Este artículo describe el sistema de clasificación de la polaridad utilizado por el equipo SINAI-word2vec en la tarea 1 del workshop TASS 2014. Nuestro sistema se basa en un método supervisado con SVM sobre la sumatoria de vectores de palabras con un modelo generado a partir de la Wikipedia en español. Nuestra solución no sigue el modelo espacio vectorial clásico ni aplica análisis sintáctico o léxico alguno. Considerando cada palabra de forma independiente representada en el espacio de 200 dimensiones de Word2Vec se consigue capturar la semántica de cada tweet y ofrecer unos resultados aceptables en la clasificación de la polaridad.
Palabras clave: Análisis de sentimientos, clasificación de la polaridad, deep-learning, Word2Vec

Abstract: In this paper it is described the participation of the SINAI-word2vec team at the task 1 on polarity classification of the TASS 2014 workshop. Our system uses supervised learning with SVM over the summatory of word vectors in a model generated from the Spanish Wikipedia. Our solution does not follow the vector space model, nor lexical or syntactic analysis is applied. Just considering each word independently represented in a 200-dimensional space of a Word2Vec model, it is possible to capture the semantics of a tweet and report good results in polarity classification.
Keywords: Sentiment analysis, polarity classification, deep learning, Word2Vec

1 Introducción

En este artículo describimos el sistema construido para participar en la tarea 1 del workshop TASS (Sentiment Analysis at global level), en su edición de 2014. Nuestra solución aplica una novedosa técnica de representación del texto basada en aprendizaje profundo, como se describe más adelante. Gracias a la aplicación del método *Word2Vec* es posible representar cada palabra en un espacio n-dimensional obtenido de los pesos asignados a nodos intermedios a partir de la relación de cada término con sus términos más próximos en un corpus de entrenamiento. Mediante el sumatorio de los vectores de las palabras que comprenden el tweet es posible obtener las características utilizadas en un proceso de aprendizaje supervisado, a partir del conjunto de entrenamiento facilitado por la organización y el algortimo SVM. Nuestros resultados son prometedores y no requieren de técnicas sofisticadas de análisis de texto.

TASS (Taller de Análisis de Sentimientos en la SEPLN) es un evento satélite del congreso SEPLN, que nace en 2012 con la finalidad de potenciar dentro de la comunidad investigadora en tecnologías del lenguaje (TLH) la investigación del tratamiento de la información subjetiva en español. En 2014 se establecen dos objetivos para este taller. Por un lado observar la evolución de los sistemas de análisis de sentimientos, y por otro lado evaluar sistemas de detección de polaridad de grano fino.

La primera tarea de TASS2014, denominada *Sentiment Analysis at global level*, consiste en el desarrollo y evaluación de sistemas que determinan la polaridad global de cada tweet del corpus general.

El resto del artículo está organizado de la siguiente forma. El capítulo 2 describe el esta-

* Esta investigación ha sido subvencionada parcialmente por el proyecto del gobierno español ATTOS (TIN2012-38536-C03-0), por la Comisión Europea bajo el Séptimo programa Marco (FP7 - 2007-2013) a través del proyecto FIRST (FP7-287607) y por el proyecto CEATIC-2013-01 de la Universidad de Jaén.

do del arte de los sistemas de clasificación de polaridad en español. En el capítulo 3 se describe el sistema desarrollado y en el capítulo 4 los experimentos realizados, los resultados obtenidos y el análisis de los mismos. Finalmente, en el último capítulo exponemos las conclusiones y el trabajo futuro.

2 Clasificación de la polaridad en español

La mayor parte de los sistemas de clasificación de polaridad están centrados en textos en inglés, y para textos en español el sistema más relevante posiblemente sea *The Spanish SO Calculator* (Brooke, Tofiloski, y Taboada, 2009), que además de resolver la polaridad de los componentes clásicos (adjetivos, sustantivos, verbos y adverbios) trabaja con modificadores como la detección de negación o los intensificadores.

Los algoritmos de aprendizaje profundo (*deep-learning* en inglés) están dando buenos resultados en tareas donde el estado del arte parecía haberse estancado (Bengio, 2009). Estas técnicas también son de aplicación en el procesamiento del lenguaje natural (Collobert y Weston, 2008), e incluso ya existen sistemas orientados al análisis de sentimientos, como el de Socher et al. (Socher et al., 2011). Aunque el aprendizaje profundo está teniendo cierta popularidad en el dominio del aprendizaje automático, y a pesar de la "novedad" con la que suele referirse a estos sistemas, sus algoritmos no son nuevos, pero sí están resurgiendo gracias a una mejora de las técnicas y la disposición de los grandes volúmenes de datos que se necesitan para su entrenamiento efectivo. Uno de los máximos expertos en estos algoritmos, Yoshua Bengio, los define como "*Un conjunto de algortimos en aprendizaje automático que intenta modelar abstracciones de alto nivel para los datos usando arquitecturas formadas por múltiples transformaciones no lineales*" (Bengio, Courville, y Vincent, 2013).

En la edición de TASS en 2012 (Saralegi Urizar y San Vicente Roncal, 2012) presentan un sistema completo de preprocesamiento de los tweets y aplican un lexicón derivado del inglés para polarizar los tweets. Sus resultados son robustos en granularidad fina (65 % de accuracy) y gruesa (71 % de accuracy). (Fernández Anta et al., 2012) presentan una comparación de diferentes técnicas de clasificación implementadas en WEKA (Hall et al., 2009). (Batista y Ribeiro, 2012) tratan la clasificación de forma binaria, y lanzan en paralelo distintos clasificadores binarios basados en regresión logística, combinando posteriormente los resultados. (Trilla y Alías, 2012) utilizan Naïve-Bayes multinomial para construir un modelo del lenguaje, obteniendo unos resultados destacables. (Martín-Wanton y Carrillo de Albornoz, 2012) utilizaron un lexicón afectivo para representar el texto como un conjunto de emociones, y resuelven el tema de la ambigüedad con un algoritmo de desambigüación (WSD) y el contexto de los términos. (Castellanos, Cigarrán, y García-Serrano, 2012) usan recuperación de información (RI), basado en divergencia del lenguaje, y concretamente Kullback-Liebler divergence, para generar modelos de polaridad. Los resultados que obtuvieron eran prometedores (quinta posición en la tarea 1), indicando que RI y modelos del lenguaje son una alternativa interesante a las propuestas clásicas en detección de polaridad. Por último, (Moreno-Ortiz y Pérez-Hernández, 2012) utilizaron un enfoque basado en el recurso léxico Sentitext, asignando una etiqueta de polaridad a cada término encontrado y calculando una puntuación o polaridad final para cada tweet.

En la edición de TASS en 2013 (Gamallo, García, y Fernández-Lanza, 2013) presentaron un sistema basado en el clasificador Naïve-Bayes, con unos buenos resultados cuando hacían clasificación binaria. (Fernández et al., 2013) presentaron un sistema con dos variantes: una versión modificada del algoritmo de ranking (RA-SR) utilizando bigramas, y una nueva propuesta basada en skipgrams. Con estas dos variantes crearon lexicones sobre sentimientos, y los utilizaron junto con aprendizaje automático (SVM) para detectar la polaridad de los tweets. Todos sus experimentos estuvieron en el top 10 de los resultados, y la combinación de ellos alcanzaron la primera posición. (Saralegi Urizar y San Vicente Roncal, 2013) utilizaron un enfoque basado en un clasificador con SVM y algún conocimiento lingüístico, obteniendo unos resultados de un 60 % de accuracy para granularidad fina y un 68 % de accuracy para gruesa. (Balahur y Perea-Ortega, 2013) utilizan aprendizaje automático y un conjunto de diferentes características. Lanzaron cuatro clasificadores en cascada para obtener la polaridad del tweet. (Vilares, Alonso,

y Gómez-Rodríguez, 2013) aplican una primera fase de normalización, seguida de un etiquetado del part-of-speech y obtienen la estructura sintáctica del tweet. En su sistema utilizan recursos para obtener las propiedades psicométricas del lenguaje, con unos resultados robustos y un buen rendimiento. (Martínez Cámara et al., 2013) optaron por una estrategia complemente no supervisada, frente a la supervisada desarrollada en 2012. Usan como recursos lingüísticos SentiWordNet, Q-WordNet y iSOL, combinando los resultados y normalizando los valores. (Villar Rodríguez et al., 2013) emplearon procesos lingüísticos avanzados, tal como la detección de negación y el tratamiento del énfasis, utilizando el recurso Freeling. (Rufo Mendo, 2013) desarrollaron un sistema basado en clasificación supervisada, y analizaron el comportamiento frente a un clasificador semi-supervisado. (Castellanos González, Cigarrán Recuero, y García Serrano, 2013) basaron su sistema en la recuperación de información, donde las clases se modelan de acuerdo a la información textual de los tweets de cada clase, y la clasificación de los tweets se usó como consulta. (Pla y Hurtado, 2013) realizaron una adecuada tokenización de los tweets, adaptando varios recursos públicos al español (como el tokenizador Tweetmotif(Krieger y Ahn, 2010)) y Freeling. En la fase de clasificación utilizaron Support Vector Machines, obteniendo unos resultados comparables a los primeros alcanzados en la competición.

Como puede verse, en general la combinación de bases de conocimiento o información externa junto con algoritmos de aprendizaje automático ofrecen las alternativas más prometedoras. Nuestro trabajo ahonda en esta linea, con un recurso adicional para la propuesta de características.

3 Descripción del sistema

Word2Vec[1] es una implementación de la arquitectura de representación de las palabras mediante vectores en el espacio continuo, basada en bolsas de palabras o n-gramas concebida por Tomas Mikolov et al. (Mikolov et al., 2013). Su capacidad para capturar la semántica de las palabras queda comprobada en su aplicabilidad a problemas como la analogía entre términos o el agrupamiento de palabras. El método consiste en proyectar las palabras a un espacio n-dimensional, cuyos pesos se determinan a partir de una estructura de red neuronal mediante un algoritmo recurrente. En la Figura 1 puede verse el modelo, basado en la topología de bolsa de palabras (CBOW), para la predicción de un término $w(t)$ en base a los términos que lo rodean (con una ventana de cinco, dos previos y dos posteriores). Existe una topología denominada *skip-gram*, muy similar, pero en la que se intenta predecir los términos acompañantes a partir de un término dado. Con estas topologías, si disponemos de un volumen de textos suficiente, esta representación puede llegar a capturar la semántica de cada palabra. El número de dimensiones (longitud de los vectores de cada palabra) puede elegirse libremente. Para el cáculo del modelo Word2Vec hemos recurrido al software indicado, creado por los propios autores del método, además de una biblioteca para Python, denominada *gensim*[2] que nos permite trabajar con el modelo generado desde este lenguaje.

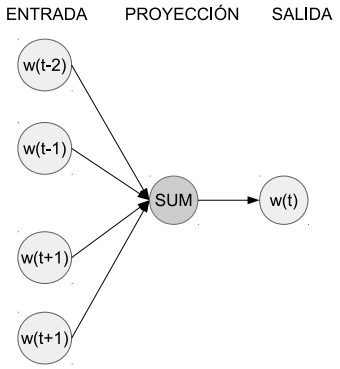

Figura 1: Modelo CBOW de Word2Vec

Tal y como se ha indicado, para obtener vectores representativos para cada palabra tenemos que generar un modelo a partir de un volumen de texto grande. Para ello hemos obtenido un volcado de Wikipedia[3] en Español de los artículos en XML, y hemos extraído el texto de los mismos. Obtenemos así unos 2,2 GB de texto plano que alimenta al programa *word2vec* con los parámetros siguientes: una ventana de 5 términos, el modelo CBOW y un número de dimensiones esperado de 200, logrando un modelo con más de 1,2 millones de palabras en su vocabulario.

Como puede verse en la Figura 2, nuestro sistema tiene dos fases de aprendizaje, una

[1] `https://code.google.com/p/word2vec/`

[2] `http://radimrehurek.com/gensim/`

[3] `http://dumps.wikimedia.org/eswiki`

en la que entrenamos el modelo de palabras haciendo uso de un volcado de la enciclopedia on-line Wikipedia, en su versión en español, como hemos indicado anteriormente, y otra en la que representamos cada tweet como la suma de los vectores de cada palabra en el tweet. Una simple normalización previa sobre el tweet es llevada a cabo, eliminando repetición de letras y poniendo todo a minúsculas. Así, el algoritmo SVM se entrena con un vector de 200 característica para cada dimensión, resultado de dicha suma. La implementación de SVM utilizada es la basada en kernel lineal proporcionada por la biblioteca Sci-kit Learn (Pedregosa et al., 2011).

Obtenemos así dos modelos: uno para los vectores de palabras según Wikipedia con Word2Vec, y otro para la clasificación de la polaridad con SVM. Esta solución es la utilizada en las dos variantes de la tarea 1 del TASS: predicción de 4 clases y predicción de 6 clases. En la Figura 3 puede verse cómo se aplican ambos modelos para obtener la predicción final.

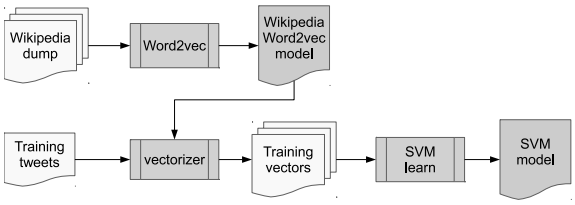

Figura 2: Flujo de datos en entrenamiento

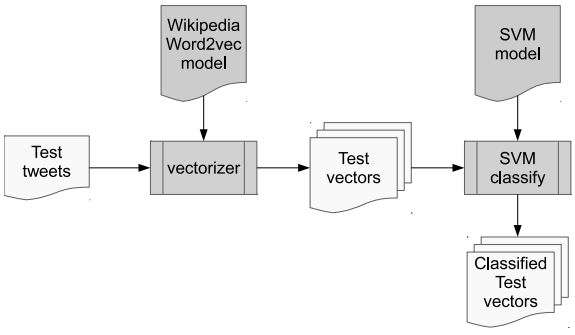

Figura 3: Flujo de datos en clasificación

4 Resultados obtenidos

Para evaluar nuestro sistema hemos realizado diversos tipos de experimentos. Estos se diferencian según varios aspectos:

- Según el nivel de sentimientos etiquetados: Aquellos que utilizan sólo tres niveles (*3l*: N, Neu, P) y los que utilizan cinco niveles (*5l*: N+, N, Neu, P, P+). En ambos casos también se etiquetan los tweets de la clase *None*.
- Según la colección de evaluación utilizada: Los organizadores pusieron a disposición de los participantes la colección completa y una colección con un número de etiquetas más homogéneo que sólo contiene 1.000 tweets. Los experimentos con esta última colección han sido nombrados como *1k*.
- Según los parámetros de SVM: Los primeros experimentos enviados a la tarea utilizaban tres etiquetas y los parámetros por defecto de SVM. Sin embargo, dichos parámetros fueron optimizados para los demás experimentos enviados.

En resumen, se enviaron 6 experimentos nombrados de la siguiente forma:

- *3l-1*: Tres niveles, colección completa y parametrización por defecto de SVM.
- *3l-1-1k*: Tres niveles, colección homogénea y parametrización por defecto de SVM.
- *3l-2*: Tres niveles, colección completa y parametrización optimizada de SVM.
- *3l-2-1k*: Tres niveles, colección homogénea y parametrización optimizada de SVM.
- *5l-1*: Cinco niveles, colección completa y parametrización optimizada de SVM.
- *5l-1-1k*: Cinco niveles, colección homogénea y parametrización optimizada de SVM.

Como se puede observar en la Tabla 1, los experimentos con mejores resultados son aquellos que utilizan tres niveles y la configuración de parámetros optimizada de SVM, llegando a alcanzar una precisión del **63 %** y un *F-score* del **77 %**. Sin embargo, al realizar una evaluación más fina, utilizando cinco niveles de sentimientos nuestros resultados descienden hasta un 51 % de precisión y un 68 % aproximadamente de *F-score*.

En el caso de utilizar una colección más homogénea con sólo 1.000 tweets, ambos experimentos con tres niveles de sentimientos

	Precission	F-score
3l-1	0,589	0,742
3l-1-1k	0,613	0,760
3l-2	0,612	0,759
3l-2-1k	**0,633**	**0,775**
5l-1	0,514	0,679
5l-1-1k	0,464	0,634

Tabla 1: Resultados obtenidos en los experimentos

mejoran sus resultados. Sin embargo, el experimento con cinco niveles no obtiene mejores resultados al utilizar la colección homogénea.

Estos datos nos indican que, aún siendo un sistema bastante sencillo, se obtienen unos resultados prometedores cuando sólo se etiqueta a tres niveles. Sin embargo nuestra clasificiación fina a cinco nivel empeora los resultados. Esto significa, que nuestra diferenciación entre sentimientos positivos y muy positivos, o negativos y muy negativos, no es suficiente.

5 Conclusiones y trabajo futuro

Este trabajo describe una novedosa aplicación de los vectores de palabras generados por el método Word2Vec a la clasificación de la polaridad, consiguiendo resultados de F-score en torno al 77 % en la competición TASS 2014, tarea 1. Estos resultados son destacables dada la simplicidad de nuestro sistema. No obstante, existen diseños experimentales que no han podido ser acometidos y que esperamos poder realizar para evaluar mejor nuestro sistema.

Los algoritmos de aprendizaje profundo prometen novedosas soluciones en el campo del procesamiento del lenguaje natural. Los resultados obtenidos con un modelo de palabras general no orientado a dominio específico alguno, ni a la tarea propia de clasificación de la polaridad, así como la no necesidad de aplicar técnicas avanzadas de análisis de texto (análisis léxico, sintáctico, resolución de anáfora, tratamiento de la negación, etc.) nos llevan a orientar nuestra investigación en una adecuación más específica de estos modelos neuronales en tareas concretas.

Es nuestra intención, por tanto, construir un modelo propio de aprendizaje profundo orientado a la clasificación de la polaridad. No dudamos que en serán numerosos los trabajos que aparecerán en este sentido en la comunidad científica, dado los buenos resultados que las técnicas de aprendizaje profundo están teniendo gracias a la disponibilidad de grandes volúmenes de datos. En cualquier caso, un diseño cuidadoso de estas redes es necesario para lograr resultados dignos del estado del arte, y ese es nuestro objetivo futuro.

Bibliografía

Balahur, Alexandra y José M. Perea-Ortega. 2013. Experiments using varying sizes and machine translated data for sentiment analysis in twitter. En *In Proc. of the TASS workshop at SEPLN 2013*.

Batista, Fernando y Ricardo Ribeiro. 2012. The l2f strategy for sentiment analysis and topic classification. En *TASS 2012 Working Notes*.

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

Bengio, Yoshua, Aaron Courville, y Pascal Vincent. 2013. Representation learning: A review and new perspectives. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8):1798–1828.

Brooke, Julian, Milan Tofiloski, y Maite Taboada. 2009. Cross-linguistic sentiment analysis: From english to spanish. En Galia Angelova Kalina Bontcheva Ruslan Mitkov Nicolas Nicolov, y Nikolai Nikolov, editores, *RANLP*, páginas 50–54. RANLP 2009 Organising Committee / ACL.

Castellanos, Angel, Juan Cigarrán, y Ana García-Serrano. 2012. Unedtass: Using information retrieval techniques for topic-based sentiment analysis through divergence models. En *TASS 2012 Working Notes*.

Castellanos González, Ángel, Juan Cigarrán Recuero, y Ana García Serrano. 2013. Considerations about textual representation for ir based tweet classification. En *In Proc. of the TASS workshop at SEPLN 2013*.

Collobert, Ronan y Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. En *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, páginas 160–167, New York, NY, USA. ACM.

Fernández, Javi, Yoan Gutiérrez, José M. Gómez, Patricio Martínez-Barco, Andrés

Montoyo, y Rafael Muñoz. 2013. Sentiment analysis of spanish tweets using a ranking algorithm and skipgrams. En *In Proc. of the TASS workshop at SEPLN 2013*.

Fernández Anta, Antonio, Philippe Morere, Luis Núñez Chiroque, y Agustín Santos. 2012. Techniques for sentiment analysis and topic detection of spanish tweets: Preliminary report. En *TASS 2012 Working Notes*.

Gamallo, Pablo, Marcos García, y Santiago Fernández-Lanza. 2013. Tass: A naive-bayes strategy for sentiment analysis on spanish tweets. En *In Proc. of the TASS workshop at SEPLN 2013*.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, y Ian H. Witten. 2009. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, Noviembre.

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

Martín-Wanton, Tamara y Jorge Carrillo de Albornoz. 2012. Uned en tass 2012: Sistema para la clasificación de la polaridad y seguimiento de temas. En *TASS 2012 Working Notes*.

Martínez Cámara, Eugenio, Miguel Ángel García Cumbreras, M. Teresa Martín Valdivia, y L. Alfonso Ureña López. 2013. Sinai-emml: Sinai-emml: Combinación de recursos lingüíticos para el análisis de la opinión en twitter. En *In Proc. of the TASS workshop at SEPLN 2013*.

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

Moreno-Ortiz, Antonio y Chantal Pérez-Hernández. 2012. Lexicon-based sentiment analysis of twitter messages in spanish. En *TASS 2012 Working Notes*.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, y others. 2011. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830.

Pla, Ferran y Lluís-F. Hurtado. 2013. Elirf-upv en tass-2013: Análisis de sentimientos en twitter. En *In Proc. of the TASS workshop at SEPLN 2013*.

Rufo Mendo, Francisco Javier. 2013. Are topic classification and sentiment analysis really different? En *In Proc. of the TASS workshop at SEPLN 2013*.

Saralegi Urizar, Xabier y Iñaki San Vicente Roncal. 2012. Tass: Detecting sentiments in spanish tweets. En *TASS 2012 Working Notes*.

Saralegi Urizar, Xabier y Iñaki San Vicente Roncal. 2013. Elhuyar at tass 2013. En *In Proc. of the TASS workshop at SEPLN 2013*.

Socher, Richard, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, y Christopher D. Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. En *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, páginas 151–161, Stroudsburg, PA, USA. Association for Computational Linguistics.

Trilla, Alexandre y Francesc Alías. 2012. Sentiment analysis of twitter messages based on multinomial naive bayes. En *TASS 2012 Working Notes*.

Vilares, David, Miguel A. Alonso, y Carlos Gómez-Rodríguez. 2013. Lys at tass 2013: Analysing spanish tweets by means of dependency parsing, semantic-oriented lexicons and psychometric word-properties. En *In Proc. of the TASS workshop at SEPLN 2013*.

Villar Rodríguez, Esther, Ana Isabel Torre Bastida, Ana García Serrano, y Marta González Rodríguez. 2013. Tecnalia-uned at tass: Uso de un enfoque lingüístico para el análisis de sentimientos. En *In Proc. of the TASS workshop at SEPLN 2013*.