TASS 2015, septiembre 2015, pp 41-46 recibido 20-07-15 revisado 24-07-15 aceptado 27-07-15

SINAI-EMMA: Vectores de Palabras para el Análisis de Opiniones en Twitter*

SINAI-EMMA: Vectors of words for Sentiment Analysis in Twitter

Eugenio Martínez Cámara, Miguel Á. García Cumbreras, M. Teresa Martín Valdivia y L. ALfonso Ureña López
Departamento de Informática
Universidad de Jaén, E-23071 - Jaén, España
{emcamara, magc, maite, laurena}@ujaen.es

Resumen: Este artículo describe el sistema de clasificación de polaridad desarrollado por el equipo SINAI-EMMA para la tarea 1 del workshop TASS 2015. Nuestro sistema construye vectores de palabras a partir de la información de opinión de 5 recursos lingüísticos. Los resultados obtenidos nos animan a seguir estudiando el aporte de los vectores de palabras a la tarea de Análisis de Opiniones.
Palabras clave: Análisis de Opiniones, Clasificación de la polaridad, recursos léxicos, Espacio Vectorial Semántico

Abstract: In this work, a polarity classification system is developed for the task 1 of workshop TASS 2015 by the SINAI-EMMA team. Our system takes advantage of 5 linguistic resources for building vectors of words. The results encourage us to continue studying the contribution of vectors of words to Sentiment Analysis.
Keywords: Sentiment Analysis, Polarity Classification, linguistic resources, Vector Space Model of Semantics

1 Introducción

TASS (Taller de Análisis de Sentimientos en la SEPLN) es un workshop satélite del congreso de la SEPLN (Sociedad Española para el Procesamiento del Lenguaje Natural), que promueve desde 2012 la evaluación experimental de sistemas de Análisis de Opiniones (AO) sobre *tweets* escritos en español.

Nuestra participación se ha limitado a la tarea 1, denominada *sentiment analysis at global level*, que tiene como objetivo la clasificación de la polaridad de un conjunto de *tweets*. La descripción de la tarea se encuentra en el resumen de la edición del año 2015 de TASS (Villena-Román et al., 2015). El sistema de este año se centra en la subtarea de de clasificación en 6 niveles de polaridad (P+, P, NEU, N, N+, NONE).

La solución que presentamos pasa por extraer características léxicas a partir de n-gramas de cada *tweet*, concretamente *unigramas* y *bigramas*, siendo estas características la polaridad de los n-gramas en distintos recursos léxicos etiquetados. Una vez obtenida la matriz de *tweets* y características utilizamos un sistema de entrenamiento para obtener un modelo léxico, y dicho modelo es el utilizado en la evaluación del subconjunto de test a etiquetar.

El resto del artículo está organizado de la siguiente forma. La siguiente sección describe los modelos de n-gramas y otros modelos del lenguaje aplicados a la detección de polaridad con *tweets*, así como trabajos relacionados. En la Sección 3 se describe el sistema que hemos desarrollado y en la Sección 4 los experimentos realizados, resultados obtenidos y análisis de los mismos. Por último, en la Sección 5 exponemos las conclusiones y el trabajo a realizar.

2 Trabajos relacionados

El Modelo de Espacio Vectorial (MEV) ha demostrado sobradamente su valía para medir la similitud entre documentos, y una muestra es su exitosa aplicación en los sistemas de recuperación de información (Manning, Raghavan, y Schütze, 2008). Por consiguiente, cabe preguntarse si es posible aplicar el mismo esquema para medir la similitud existente entre palabras. Esa pregunta la contestaron afirmativamente Deerwester et al. (1990) al proponer un MEV en el

* Este trabajo ha sido financiado parcialmente por el Fondo Europeo de Desarrollo Regional (FEDER), el proyecto ATTOS del Gobierno de España (TIN2012-38536-C03-0) y el proyecto AORESCU de la Junta de Andalucía (P11-TIC-7684 MO).

que cada palabra estaba caracterizada por un vector, el cual representaba a la palabra en función a su coocurrencia con otras. Al igual que el MEV sobre documentos, el MEV sobre palabras también ha reportado buenos resultados en desambiguación, reconocimiento de entidades, análisis morfológico (*part of speech tagging*) y recuperación de información (Turney y Pantel, 2010; Collobert y Weston, 2008; Turian, Ratinov, y Bengio, 2010).

En la bibliografía relacionada con AO existen trabajos que postulan que las palabras (*unigramas*) son más efectivos para representar la información de opinión (Pang, Lee, y Vaithyanathan, 2002; Martínez-Cámara et al., 2011), y otros en los que se prefiere el uso de n-gramas (Dave, Lawrence, y Pennock, 2003; Kouloumpis, Wilson, y Moore, 2011). La orientación de la opinión de un documento es muy probable que no sea la suma de las polaridades de las palabras individuales que lo componen, sino la combinación de las expresiones que aparecen en el mismo. La palabra "bueno" tiene un significado positivo, pero la expresión "no muy bueno" transmite un significado negativo. Por tanto, en el caso del AO puede que sea más recomendable utilizar n-gramas como unidad de representación de información.

En el presente trabajo tratamos de incrustar al MEV tradicional, una representación vectorial de cada n-grama, dando lugar a un MEV semántico (Turney y Pantel, 2010). Este tipo de modelo tiene una mayor capacidad de representación de la información subyacente en la opinión, al combinar diferentes grados de n-gramas, así como de caracterizar cada n-grama por un conjunto de características. Éste es el fundamento de los nuevos métodos que se están desarrollando para la clasificación de la polaridad. Uno de los primeros trabajos de este nuevo enfoque es (Maas et al., 2011), en el que se consigue que palabras con un significado similar estén representadas por vectores similares. El proceso de generación de los vectores es no supervisado, pero para una correcta clasificación de la polaridad, el método que proponen los autores requiere de información de opinión, de manera que tienen que utilizar un corpus con etiquetas de opinión para conseguir que los vectores asociados a las palabras representen la orientación semántica de la misma.

Socher et al. (2011) tratan de representar la opinión de un conjunto de documentos a través de un MEV a nivel de n-grama, siguiendo en este caso, un modelo de autocodificadores recursivos (*recursive autoencoders*). Bespalov et al. (2012) tratan tener en cuenta en el MEV semántico que diseñan la posición de cada n-grama en el documento, dado que sostienen que la polaridad de un documento depende de la posición en la que se encuentren los n-gramas. El trabajo (Tang et al., 2014) es otro ejemplo de representación de n-gramas por medio de vectores, tratando los autores de insertar información de opinión en dichos vectores mediante el uso de tres redes neuronales. A diferencia de los trabajos anteriores, el de Tang et al. (2014) se centra en la clasificación de la polaridad de *tweets*.

3 Sistema

El sistema que se ha desarrollado para la cuarta edición de TASS trata de trasladar el concepto de MEV semántico a la clasificación de la polaridad de *tweets* en español. En nuestro caso, no se ha pretendido representar a cada palabra en función de su coocurrencia con otras, o teniendo en cuenta su posición en el texto, sino por su valor de polaridad en varios recursos léxicos de opinión: iSOL (Molina-González et al., 2013), SentiWordNet (Baccianella, Esuli, y Sebastiani, 2010), Q-WordNet (Agerri y García-Serrano, 2010), SEL (Rangel, Sidorov, y Suárez-Guerra, 2014) y ML-Senticon (Cruz et al., 2014). Con este esquema de representación, lo que se está haciendo es incrustar en el MEV los distintos puntos de vista que tienen diferentes recursos léxicos de opinión sobre una misma palabra, enriqueciendo de esta forma la representación de la información de opinión que contienen cada una de las palabras.

A continuación se van a describir someramente los recursos lingüísticos que se han utilizado, así como se detallará el proceso de generación de los vectores.

3.1 iSOL

iSOL es una lista de palabras indicadoras de opinión en español. La lista está formada por 8135 palabras, de las cuales 5626 son negativas y 2509 son positivas.

Al tratarse de una lista de palabras de opinión, la única información que aporta es si un término es positivo o negativo, por lo que proporciona dos características binarias al vector de cada palabra.

3.2 SentiWordNet

SentiWordNet es un lista de sentidos de opinión que se construyó siguiendo un enfoque basado en diccionario (Liu, 2012). El diccionario que se empleó en su desarrollo fue WordNet (Miller, 1995). Por tanto, SentiWordNet asocia a cada sentido (*synset*) de WordNet tres valores probabilidad de pertenencia a tres niveles de polaridad: Positivo, Negativo y Neutro.

Cada palabra puede tener asociada varios *synsets*, por lo que se necesita de una función de agregación para obtener un valor de polaridad único para cada uno de los tres niveles de opinión. Al igual que Denecke (2008), se empleó como función de agregación la media aritmética. Por tanto, se calculó la media aritmética de cada nivel de polaridad, obteniéndose de esta manera 3 características, cada una correspondiente a un nivel de polaridad.

3.3 Q-WordNet

Q-WordNet, al igual que SentiWordNet, está basado en WordNet para construir un recurso para la tarea de AO. Los autores de Q-WordNet, al contrario que los de SentiWordNet, consideran la polaridad como una propiedad cualitativa, de manera que una palabra sólo puede ser positiva o negativa. Por consiguiente, Q-WordNet es un recurso de opinión que asocia a cada *synset* de WordNet el valor Positivo o Negativo. El uso de Q-WordNet permite la adición de 2 nuevas características al vector asociado a cada palabra.

3.4 SEL

SEL es una lista de 2036 palabras clasificadas en 6 estados de ánimo diferentes: alegría, enfado, miedo, tristeza, sorpresa y disgusto. Las palabras tienen asociado un valor de probabilidad, que los autores llaman PFA, de pertenencia a una de las 6 categorías.

Para su integración en el sistema, se han transformado las 6 categorías emocionales en dos valores de polaridad, de forma que las categorías alegría y sorpresa se han tomado como positivas, y las clases enfado, miedo, tristeza y disgusto como negativas. Además, se aplicó un filtro a la lista de palabra con el fin de sólo utilizar aquellas que tuvieran un valor de PFA superior a 0,2. Por tanto, tras convertirse SEL a una lista binaria de opinión, se pudieraon generar dos nuevas características binarias de polaridad.

3.5 ML-SentiCon

ML-SentiCon es un recurso de opinión a nivel de lema. Los 26495 que conforman ML-SentiCon se encuentran estratificados en 8 capas. Cada una de las capas representa un nivel de exactitud de pertenencia a la clase positivo o negativo, de manera que existe una mayor probabilidad, de que el significado de opinión que transmiten los lemas de la capa 0 coincida con la clase de opinión que se le ha asignado que los de la capa 7. Al estar los lemas catalogados como positivos o negativos, este recurso también genera dos nuevas características, pero en lugar de ser binarias, su valor se corresponde con la puntuación de polaridad que tiene cada lema en el recurso. En el caso de que el lema sea negativo, su puntuación se multiplica por -1.

4 Clasificación

La preparación de la clasificación consistió en el procesamiento adecuado de los *tweets* para la generación del MEV semántico de *unigramas* y bigramas. La preparación comenzó con la *tokenización* de los *tweets*; aplicación de un proceso de normalización léxica que consistió principalmente en la corrección ortográfica de los tokens; lematización y aplicación de un análisis morfológico.

A continuación se construyeron los vectores de características asociados a cada *unigrama*. En función de la naturaleza de cada recurso, la búsqueda fue distinta. En el caso de iSOL se utilizó como consulta la forma original de cada *unigrama*, a excepción de los *unigramas* que se corresponden con verbos, con los que se usó su lema, es decir, el infinitivo del verbo.

Para SentiWordNet y Q-WordNet la consulta no fue directa porque ambos son dos recursos para inglés. Por tanto, se precisó emplear un recurso que permitiera a partir de palabras en español obtener el *synset* que le corresponde en WordNet. Ese recurso fue la versión en español de WordNet de Multilingual Central Repository (MCR) (Atserias et al., 2004). Una vez que se tenían de MCR los *synsets* asociados, se consultaba SentiWordNet y Q-WordNet para obtener sus correspondientes valores de polaridad.

SEL y ML-Senticon tuvieron un tratamiento similar, dado que la consulta se realizaba utilizando el lema de cada *unigrama*.

Antes de continuar con la descripción de la generación de los *bigramas*, debe precisarse que únicamente se construyeron los vectores de los *unigramas* que al menos estaban recogidos en uno de los cinco recursos lingüísticos que se han considerado en la experimentación. Tras la eliminación de los *unigramas* que no transmiten opinión, el texto del *tweet* se quedaba reducido a dichos *unigramas*. Con esos *unigramas* se desarrolló un proceso de generación de *bigramas*, cuyos vectores se corresponden con la suma vectorial de los vectores de los *unigramas* que constituyen el *bigrama*. La longitud de los vectores de características de opinión de tanto los *unigramas* como de los *bigramas* es de 11 características.

Una vez generado el MEV semántico, es momento de generar el modelo de entrenamiento. Para ello se eligió el algoritmo SVM con *kernel* lineal, ya que en experimentaciones previas ha demostrado su valía para la tarea de AO sobre *tweets*. La implementación que se utilizó fue la de la librería scikit-learn[1] de Python.

5 Resultados

Primeramente, se desarrolló una evaluación con el conjunto de entrenamiento, para evaluar si el esquema descrito anteriormente podría tener éxito.

La evaluación consistió en aplicar un marco de trabajo basado en validación cruzada de 10 particiones (*ten fold cross-validation*). Las medidas de evaluación que se utilizaron fueron la comunes en tareas de clasificación de textos, es decir, Precisión, *Recall*, F1 y *Accuracy*. La tarea a la que se enfrenta el sistema es la de clasificación en seis niveles de de intensidad de opinión, de manera que el uso de la definición clásica de Precisión, *Recall* y F1 no es correcto, debido principalmente a que el resultado estaría sesgado por la clase predominante del conjunto de datos que se está utilizando. Por consiguiente, se decidió usar sus correspondientes macromedidas (*macro-averings*).

Se llevaron a cabo dos evaluaciones, que se diferencian en la manera de aprovechar la información de opinión que aporta el recurso iSOL. En una primera ejecución (SINAI-EMMA_iSOLOriginal), solamente se buscaban en iSOL las formas originales de las palabras convertidas en minúsculas. Si se comprueba el léxico recogido en iSOL se puede comprobar que en él se recogen vocablos, y que uno de los puntales de su éxito es que incluye las derivaciones de género y número propias del español. En cambio, en lo que se refiere a los verbos, no incluye las conjugaciones de los mismos, estando sólo presente el infinitivo. Por tanto, en esta primera ejecución se estaba perdiendo la cobertura de las distintas formas verbales, ya que sólo se podían reconocer aquellas que en el *tweet* aparecieran en infinitivo.

La segunda configuración (SINAI-EMMA_iSOLLema) tuvo en cuenta esa peculiaridad de iSOL, y en el caso de que el término a buscar fuera un verbo, se empleó su lema, es decir su infinitivo. De esta manera se consiguió que un mayor número de vocablos fueran considerados como portadores de opinión. En la Tabla 1 se muestran los resultados alcanzados por estas dos configuraciones.

No es difícil dar una explicación a la ligera mejoría de la configuración SINAI-EMMA_iSOLLema, dado que posibilita aumentar la cobertura del lenguaje por parte de iSOL. Por tanto, la configuración que se utilizó para la clasificación de los *tweets* del subcorpus de test es SINAI-EMMA_iSOLLema. La Tabla 2 recoge los resultados oficiales obtenidos por el sistema, así como el mejor resultado y la media de los resultados del resto de sistemas presentados.

Los resultados evidencian que el sistema se encuentra en la media de los presentados en la edición 2015 de TASS, lo cual por un lado es satisfactorio, ya que el estudio que se ha iniciado en el ámbito de los MEV semánticos reporta, por ahora, unos resultados similares al resto de enfoques que se han presentado, pero por otro, pone de manifiesto que todavía quedan muchos elementos que analizar para seguir avanzando en la resolución del problema de la clasificación de la polaridad de *tweets* en español.

La organización del taller proporciona también una evaluación con un conjunto reducido de 1000 *tweets* etiquetados a mano. La Tabla 3 recoge los resultados obtenidos sobre ese conjunto de datos.

Como se puede apreciar en la Tabla 3, las diferencias entre el sistema presentado y el mejor se acortan, a su vez que el sistema consigue unos resultados superiores a la media según la Macro-Precisión, el Macro-*Recall*

[1] `http://scikit-learn.org/`

Configuración	**Macro-P**	**Macro-R**	**Macro-F1**	***Accuracy***
SINAI-EMMA_iSOLOriginal	36,55 %	36,58 %	35,99 %	40,91 %
SINAI-EMMA_iSOLLema	36,83 %	36,74 %	36,02 %	41,31 %

Tabla 1: Evaluación del sistema de clasificación con el conjunto de entrenamiento.

Configuración	**Macro-P**	**Macro-R**	**Macro-F1**	***Accuracy***
SINAI-EMMA	40,4 %	45,8 %	43,3 %	50,02 %
Mejor	53,1 %	46,5 %	49,6 %	67,3 %
Media	40,9 %	42,2 %	40,9 %	52,7 %

Tabla 2: Resultados de la evaluación oficial.

y el Macro-F1. Ésto indica que el sistema SINAI-EMMA tiene un comportamiento más estable que el resto, porque su rendimiento no experimenta un variación muy acusada ante la modificación del etiquetado del conjunto de evaluación, lo cual es un punto positivo en nuestra investigación.

6 Conclusiones y trabajo a realizar

La principal conclusión a la que se ha llegado es la idoneidad de la representación de la información de opinión mediante un MEV semántico. Además, características de los *unigramas* y *bigramas* se han construido a partir de recursos lingüísticos de opinión. Esto es una primera tentativa a la aplicación de MEV semánticos a la tarea de AO, y los resultados obtenidos nos animan a seguir investigando en esta línea.

El trabajo futuro va a estar dirigido a mejorar el esquema de representación de información, así como en aumentar la información de opinión que representa a cada palabra. Para ello, se va a realizar un estudio del nivel de cobertura de vocabulario que tiene el esquema de representación; se va a tratar de introducir información de la distribución de los *unigramas* y *bigramas* en el corpus por medio del uso de la frecuencia de los mismos en cada documento y en el corpus en general; se va a estudiar la incorporación de un mayor número de recursos de opinión; se va a analizar el aporte para el AO de la incrustación de información de coocurrencia de cada *unigrama* y *bigrama* tomando como referencia un corpus representativo del español; y por último se va a estudiar la utilización de información sintáctica mediante la consideración del efecto de la negación y de los intensificadores.

Bibliografía

Agerri, Rodrigo y Ana García-Serrano. 2010. Q-Wordnet: Extracting polarity from wordnet senses. En *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta; ELRA, may. European Language Resources Association. 19-21.

Atserias, J., L. Villarejo, G. Rigau, E. Agirre, J. Carroll, B. Magnini, y P. Vossen. 2004. The meaning multilingual central repository. En *GWC 2012 6th International Global Wordnet Conference*. Brno: Masaryk University.

Baccianella, S., A. Esuli, y F. Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. En *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, páginas 2200–2204, Valletta, Malta.

Bespalov, Dmitriy, Yanjun Qi, Bing Bai, y Ali Shokoufandeh. 2012. Sentiment classification with supervised sequence embedding. En *Machine Learning and Knowledge Discovery in Databases*, volumen 7523 de *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, páginas 159–174.

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.

Cruz, F. L., J. A. Troyano, B. Pontes, y F. J. Ortega. 2014. Ml-senticon: Un lexicón

Configuración	**Macro-P**	**Macro-R**	**Macro-F1**	***Accuracy***
SINAI-EMMA	36,6 %	38,0 %	37,3 %	41,1 %
Mejor	44,1 %	45,9 %	45,0 %	51,6 %
Media	36,3 %	37,28 %	36,68 %	41.35 %

Tabla 3: Resultados oficiales sobre el subcorpus de 1000 *tweets*.

multilingüe de polaridades semánticas a nivel de lemas. *Procesamiento del Lenguaje Natural*, 53(0):113–120.

Dave, K., S. Lawrence, y D. M. Pennock. 2003. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. En *Proceedings of the 12th international conference on World Wide Web*, WWW '03, páginas 519–528, New York, NY, USA. ACM.

Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer, y R. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Kouloumpis, Efthymios, Theresa Wilson, y Johanna Moore. 2011. Twitter sentiment analysis: The good the bad and the omg!

Liu, B. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Maas, A. L., R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, y C. Potts. 2011. Learning word vectors for sentiment analysis. En *Proceedings of the 49th Annual Meeting of the ACL: Human Language Technologies*, volumen 1 de *HLT '11*, páginas 142–150. ACL.

Manning, Christopher D., Prabhakar Raghavan, y Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Martínez-Cámara, Eugenio, M. Teresa Martín-Valdivia, José M. Perea-Ortega, y L. Alfonso Ureña-López. 2011. Técnicas de clasificación de opiniones aplicadas a un corpus en español. *Procesamiento del Lenguaje Natural*, 47:163–170.

Miller, George A. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, Noviembre.

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 Syst. Appl.*, 40(18):7250–7257.

Pang, Bo, Lillian Lee, y Shivakumar Vaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. En *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, volumen 10 de *EMNLP '02*, páginas 79–86. ACL.

Rangel, I. D., G. Sidorov, y S. Suárez-Guerra. 2014. Creación y evaluación de un diccionario marcado con emociones y ponderado para el español. *Onomázein*, 1(29):31–46.

Socher, R., J. Pennington, E. H. Huang, A. Y. Ng, y C. D. Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. En *Proceedings of the Conference on EMNLP*, EMNLP '11, páginas 151–161. ACL.

Tang, Duyu, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, y Bing Qin. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. En *Proceedings of the 52nd Annual Meeting of the ACL*, volumen 1, páginas 1555–1565, Baltimore, Maryland, June. ACL.

Turian, Joseph, Lev Ratinov, y Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. En *Proceedings of the 48th Annual Meeting of the ACL*, ACL '10, páginas 384–394. ACL.

Turney, Peter D. y Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *J. Artif. Int. Res.*, 37(1):141–188, Enero.

Villena-Román, Julio, Janine García-Morera, Miguel A. García-Cumbreras, Eugenio Martínez-Cámara, M. Teresa Martín-Valdivia, y L. Alfonso Ureña López. 2015. Overview of tass 2015. En *TASS 2015: Workshop on Sentiment Analysis at SEPLN*.