Análisis de Sentimientos

Eugenio Martínez Cámara, MªTeresa Martín Valdivia, L. Alfonso Ureña

SINAI - Sistemas Inteligentes de Acceso a la Información
Departamento de Informática
Universidad de Jaén
{emcamara,maite,laurena}@ujaen.es

Entre una de las múltiples tareas de las que se ocupa el PLN se encuentra la clasificación de textos, que consiste en la asignación de un conjunto de categorías a una colección de documentos, resolviéndose de esta forma la clasificación objetiva de documentos.

Existe una gran cantidad de textos en el que el contenido subjetivo es lo más relevante, y cuyo procesamiento no debería limitarse a aplicar únicamente las técnicas de la clasificación de documentos. Ante esta necesidad de clasificar la orientación, o la opinión que se expresan en los documentos, surge la área análisis de sentimientos (AS), o también denominada en la bibliografía como minería de opiniones, o en inglés, sentiment analysis u opinion mining.

El análisis de sentimientos trata de clasificar los documentos en función de la polaridad de la opinión que expresa su autor. Esta nueva área que combina PLN y minería de textos, incluye una gran cantidad de tareas que han sido tratadas en mayor o menor medida [3]. Existen principalmente dos formas distintas de enfrentarse a este problema: aplicando aprendizaje automático [2] o aplicando un enfoque semántico [1]. Dos son las aplicaciones más importantes: determinar la polaridad de las opiniones a nivel de documento, frase o característica, y determinar si un documento contiene opiniones.

Existen muchos trabajos en el campo del análisis de sentimientos, habiéndose aplicado en multitud de dominios, pero la mayor parte de ellos han sido realizados sobre corpus de documentos en inglés. La investigación en AS en español es reducida, siendo la más destacable la que está llevando a cabo el grupo ITALICA de la universidad de Sevilla. El grupo SINAI de la universidad de Jaén también lleva algún tiempo trabajando en esta temática pero centrándose hasta hace unos meses solamente en lengua árabe [5].

Al final del año pasado hicimos una primera aproximación al estudio del análisis de opiniones en español. Se eligió el enfoque de aprendizaje automático propuesto por Pang en [2], por su sencillez, por los buenos resultados que ha dado en trabajos anteriores, y para poder comparar los resultados con los obtenidos por Cruz y otros [4], ya que ellos utilizan el enfoque semántico propuesto por Turney [1]. Se aplicaron diversos algoritmos de clasificación, dos de los que se suelen utilizar en AS, SVM y Naïve Bayes, y otros tres, BBR (regresión logística bayesiana), KNN y C4.5, para determinar que algoritmo se comporta mejor para este problema. Para ello se utilizó un corpus de críticas de cine en español [4], compuesto por 3.878 críticas recogidas de la web *muchocine*[1]. Las críticas están

[1] http://www.muchocine.net

puntuadas en un rango de 1 a 5, significando el 1 una película muy mala, y el 5 una película muy buena. Solamente se utilizaron para la experimentación las puntuadas con 1, 2, 4, 5, eliminado las que tienen un 3 debido a su carácter neutro. Los mejores resultados se obtuvieron aplicando regresión logística bayesiana, es decir, con el algoritmo BBR, el cual no había sido utilizado antes en trabajos de análisis de sentimientos. También hay que destacar el hecho de que se mejoraron considerablemente los resultados obtenidos por Cruz y otros [4] sobre el mismo conjunto de datos.

En los últimos años ha aumentado exponencialmente el uso de las redes sociales en todo el mundo. En todas ellas los usuarios vierten opiniones de cualquier tipo y sobre cualquier tema. Dentro del conjunto de las redes sociales las de microblogging, en concreto *Twitter*[2], no han parado de crecer tanto en el número de usuarios como en el tráfico de datos. En España, en los últimos meses el uso de *Twitter* ha aumentado de forma muy considerable. Como se puede ver, existe una gran cantidad de información que se genera en internet y no se debe dejar de aprovechar la oportunidad de procesarla. Desde finales 2009 están apareciendo trabajos muy interesantes en los que se aplica AS a *Twitter*, pero todos ellos en inglés [6], [7] o [8]. Entre la infinidad de aplicaciones que puede tener aplicar AS a *Twitter* se encuentra el poder determinar la opinión de los usuarios sobre una marca comercial, sobre el último producto presentado por una compañía, sobre la campaña electoral de un partido político, o estimar la intención de voto de un determinado partido político.

Nuestro trabajo ahora mismo se centra en la aplicación de técnicas de análisis de sentimientos a información en español recogida de *Twitter*, sin olvidar la continuación de nuestro primer trabajo de análisis de opiniones en textos en español.

Referencias

1. Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL) pp. 417–424. ACL. Morristown, NJ, USA.
2. Pang, B., Lee, L.,Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) pp. 79–86. Association for Computational Linguistics.
3. Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. Foundation and Trends in Information Retrieval, vol. 2, nos. 1-2, pp. 1–135.
4. Cruz, F.L, Troyano, J.A, Enriquez, F., Ortega J. (2008). Clasificación de documentos basada en la opinión: experimentos con un corpus de críticas de cine en español. In: Procesamiento de Lenguaje Natural, vol 41.
5. M. Saleh, A. Montejo, M.T. Martín, L.A. Ureña. (2009). Prediction of Customer Ratings on a New Corpus for Opinion Mining. Proceedings Working Notes for the WOMSA 2009 Workshop. Sevilla.

[2] http://twitter.com

6. A. Go, R. Bhayani, L. Huang. (2009). Twitter Sentiment Classification using Distant Supervision. CS224N Project Report. Stanford.
7. O'Connor, B. Balasubramanyan R. Routledge, B. Smith, N. (2010). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. International AAAI Conference on Weblogs and Social Media, North America.
8. Tumasjan, A. Sprenget, T. Sandner, P. Welpe, I. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. International AAAI Conference on Weblogs and Social Media, North America.