Procesamiento de Lenguaje Natural, Revista nº 56, marzo de 2016, pp 103-106 recibido 20-11-2015 revisado 16-12-2015 aceptado 06-02-2016

Análisis de Opiniones en Español*

Sentiment Analysis in Spanish

Eugenio Martínez Cámara
Departamento de Informática - Universidad de Jaén
Campus Las Lagunillas, E,-23071, Jaén, España
emcamara@ujaen.es

Resumen: Tesis doctoral elaborada por E. Martínez Cámara en la Universidad de Jaén bajo la dirección de los doctores D. L. Alfonso Ureña López y Dª. M. Teresa Martín Valdivia. La defensa tuvo lugar el 26 de octubre de 2015 en Jaén ante el tribunal formado por la doctora Dª. María Teresa Taboada Gómez de la Universidad Simon Fraser (Canadá) como presidenta, por el doctor D. José Manuel Perea Ortega de la Universidad de Extremadura (España) como secretario y por la doctora Dª. Alexandra Balahur Dobrescu del *Joint Research Centre* (Italia) de la Comisión Europea como vocal. La tesis obtuvo la mención Internacional y logró una calificación de Sobresaliente Cum Laude.
Palabras clave: Análisis de Opiniones, aprendizaje supervisado, aprendizaje no supervisado, combinación de clasificadores, generación de recursos

Abstract: Ph.D. thesis written by Eugenio Martínez Cámara at the University of Jaén under the supervision of the Ph.D. L. Alfonso Ureña López and the Ph.D. M. Teresa Martín Valdivia. The author was examined on 26st October 2015 by a pannel composed by the Ph.D. María Teresa Taboada Gómez from the Simon Fraser University (Canada) as president of the pannel, the Ph.D. José Manuel Perea Ortega from the University of Extremadura (Spain) as secretary of the pannel and the Ph.D. Alexandra Balahur Dobrescu from the Joint Research Centre (Italy) of the European Comission as a panel member. The Ph.D. was awared Summa cum laude and it obtained the International mention.
Keywords: Sentiment Analysis, supervised learning, unsupervised learning, ensemble classifiers, linguistic resources generation

1 Introducción

En nuestro día a día, las personas nos enfrentamos a una cantidad importante de decisiones que tomar, y esas decisiones pueden requerir de un menor o mayor grado de reflexión. Las simples, es muy probable que no requieran de un detenimiento excesivo, pero las complejas pueden que necesiten de la incorporación de información externa. El proceso por el cual se inserta información externa se conoce como "petición de opinión". La "petición de opinión" suele realizarse en primer lugar a personas de nuestro entorno de confianza. Pero, en ocasiones la información u opinión que requerimos no puede ser facilitada por nuestro entorno de confianza porque está circunscrita a una temática especializada. En estos casos se debe acudir a opiniones especializadas sobre la cuestión que nos interesa. Tradicionalmente el método por el que se llevaba a cabo un proceso de "petición de opinión" a nuestro círculo de confianza era la comunicación "boca a boca", mientras que para ilustrarnos con opiniones especializadas había que acudir a la prensa o publicaciones de una temática concreta.

A finales del siglo XX la Web comenzó paulatinamente a revolucionar la sociedad. La primera etapa de esa revolución fue la simplificación del acceso a información, a la cual anteriormente era verdaderamente complicado llegar a ella. La segunda etapa de esa revolución está magníficamente representada por el hito del advenimiento de la Web 2.0. El concepto de Web 2.0 significa la ruptura de las barreras existentes entre pro-

* Este trabajo de investigación ha sido parcialmente financiado por el Fondo Europeo de Desarrollo Regional (FEDER), el proyecto FIRST FP7-287607 del Séptimo Programa Marco para el Desarrollo de la Investigación y la Tecnología de la Comisión Europea; el proyecto ATTOS TIN2012-38536-C03-0 del Ministerio de Economía y Competitividad y el proyecto AROESCU P11-TIC-7684 MO de Excelencia de la Junta de Andalucía.

ISSN 1135-5948

ductores y consumidores de información, lo cual supuso que cualquier persona, incluso sin disponer de conocimientos de informática, pudiera de una manera sencilla publicar cualquier tipo de contenido.

La Web 2.0 auspició la llegada de nuevas plataformas en las que se ponían en contacto, de una manera casi natural, a productores y consumidores de información. Si se analiza la información que se publica en las distintas plataformas propias de la Web 2.0 y sobre todo la que fluye a través de las redes sociales, se puede comprobar que la información subjetiva o las opiniones son un porcentaje importante del total. Que el nivel de presencia de opiniones en la Web sea elevado no debe causar extrañeza, dado que la petición de opinión y la expresión de la misma es propia de la condición humana.

Por tanto, existen una gran cantidad de información subjetiva u opiniones en Internet, que debido a su enorme volumen, si no se facilita su acceso, sería complicado para una persona encontrar aquellos puntos de vista que pueden ser positivos para su proceso de toma de decisiones. Como consecuencia, se hace necesario el desarrollo de sistemas que permitan la identificación de opiniones, así como la determinación automática de la orientación de las mismas.

Pero el conocimiento de las opiniones de otras personas no sólo es de interés para la persona que está inmersa en un proceso de tomas de decisiones, sino también a todos aquellos entes sobre los cuales se está opinando. Es evidente que a una institución pública, a un partido político o a una empresa le interesa conocer lo que se está opinando sobre ellos, porque se trata de una herramienta muy valiosa de obtener información. Por ende, no sólo es necesario facilitar el acceso a opiniones a los usuarios ávidos de conocer experiencias similares, sino también a todos aquellos entes sobre los cuales se está opinando.

Entre las diversas tareas relacionadas con el Procesamiento del Lenguaje Natural, el Análisis de Opiniones (AO) es la responsable del tratamiento automático de opiniones. Tradicionalmente se ha definido formalmente el AO como la tarea que se encarga del tratamiento computacional de opiniones, sentimientos y de la subjetividad en los textos. Actualmente se emplea una definición más completa, la cual indica que el AO se corresponde con el conjunto de técnicas computacionales para la extracción, clasificación, comprensión y evaluación de opiniones expresadas en fuentes publicadas en Internet, comentarios en portales web y en otros contenidos generados por usuarios.

Si se estudia el estado actual de la investigación en AO, se puede comprobar que la mayor parte de la invetigación está orientada al tratamiento del inglés. Pero la información que fluye constantemente por Internet no solamente etá expresada en inglés, sino también en otras muchas lenguas. El español es la segunda lengua materna del mundo por número de hablantes y la tercera lengua más utilizada en Internet, de manera que parece perentoria la necesidad de estudiar métodos de tratamiento automático de opiniones expresadas en español.

Resumiendo lo que se ha indicado en los párrafos anteriores, las motivaciones sobre las que se sustenta la tesis son: la necesidad de desarrollar métodos que posibiliten la determinación automática de la orientación de la opinión, con la intención de facilitar el acceso a la información de opinión a cualquier tipo de personas, así como a todos aquellos entes interesados en conocer que se está opinando sobre ellos; y la necesidad de adaptar dichos métodos al tratamiento de opiniones expresadas en español.

2 Organización de la memoria

La tesis expone un amplio estudio de diversas técnicas de clasificación de la polaridad de opiniones, en el que se ha intentado cubrir los diferentes enfoques existentes en la clasificación de textos. También no se ha querido dejar a un lado la generación de recursos, debido principalmente a su importancia para insertar conocimiento a los sistemas de clasificación de la polaridad. Asimismo se debe indicar, que las técnicas estudiadas se han evaluado teniendo en cuenta textos largos, o provenientes de comentarios publicados en la Web, y textos cortos, principalmente aquellos que se han publicado en la red social Twitter.

La amplitud del estudio que se ha llevado a cabo se puede comprobar a través de la propia estructura de la tesis, la cual se va a desgranar en los siguientes párrafos.

El Capítulo 1 se centra en desarrollar ampliamente la necesidad de estudiar el tratamiento automático de opiniones, atendiendo a la creciente cantidad de información presente en Internet relacionada con estados perso-

nales, a la demanda de tanto la sociedad como la industria de procesar esa ingente cantidad de información y la necesidad de desarrollar métodos para el procesamiento de opiniones escritas en español.

El Capítulo 2 trata de definir la tarea del AO y situarla en el contexto del PLN. Además, en dicho capítulo se presenta el estado en el que se encuentra actualmente la investigación relacionada con el AO.

El Capítulo 3 tiene una misión propedéutica en relación a la investigación que se va a describir en los siguientes capítulos. La evaluación de los distintos métodos se ha llevado a cabo utilizando dos tipos de documentos: textos largos y textos cortos. En el Capítulo 3 se definen estos dos tipos de documentos y se comparan sus principales características. Asimismo, se definen las medidas de evaluación que se van a utilizar para medir la bondad de los distintos métodos de inferencia.

El estudio de la opinión se puede desarrollar en tres niveles distintos, dependiendo de la granulidad del análisis. Dichos niveles son: documento, oración y aspecto. Por último, en el Capítulo 3 se indica que el nivel de análisis que se ha aplicado en cada uno de los experimentos ha sido el de documento.

El Capítulo 4 se centra en la descripción de todos los sistemas que se han desarrollado siguiendo una metodología de aprendizaje supervisado. Para las experimentaciones relacionadas con textos largos se emplearon corpus de opiniones disponibles, pero para las correspondientes con textos cortos fue necesario la generación de un corpus de *tweets* etiquetados en función de la opinión que expresan. La descripción de dicho corpus, el cual se denomina COST (*Corpus of Spanish Tweets*), se encuentra también recogida en el Capítulo 4.

El Capítulo 5 se centra en detallar la experimentación que se ha realizado siguiendo un enfoque no supervisado. En este Capítulo se presenta un método modular e independiente del idioma para la clasificación de la polaridad. La principal novedad del método es que trata de representar cada uno de los términos que aparecen en los textos como vectores de conceptos relacionados en función de su significado en el contexto en el que se encuentran.

El Capítulo 6 está dedicado por un lado a la presentación de los recursos lingüísticos para la invetigación en AO que se han elaborado durante la tesis, y por otro, se describe las experimentaciones que se han llevado a cabo en el ámbito de la adaptación al dominio.

El Capítulo 7 recoge las experimentaciones que se han desarrollado siguiendo un enfoque de combinación de clasificadores. El español no cuenta con una nutrida cantidad de recursos lingüísticos que se puedan utilizar en sistemas de clasificación de la polaridad. Por este motivo, se ha tratado de superar esa barrera mediante la definición de sistemas que combinan clasificadores especializados en el tratamiento de opiniones en inglés, y clasificadores de opiniones en español.

El Capítulo 8 destaca las conclusiones a las que se han llegado durante la elaboración de la tesis, así como presenta las intenciones relacionadas con los siguientes trabajos que se van a emprender.

3 Contribuciones más relevantes

De cada uno de los capítulos se pueden extraer importantes contribuciones. Comenzando con las relacionadas con la generación de recursos, debe destacarse que la tesis ha dado como resultado tres recursos, los cuales son: el corpus de *tweets* COST[1] (Martínez-Cámara et al., 2015), la lista de palabras de opinión iSOL[2] (Molina-González et al., 2013) y el corpus de opiniones en el dominio del alojamiento hotelero COAH (*Corpus of Andalusian Hotels*) [3] (Molina-González et al., 2014).

En relación a la experimentación supervisada, se llegó a la conclusión de que los textos a los que se han venido a llamar largos requieren de un procesamiento distinto al que necesitan los textos que se han venido a llamar cortos. Cuando los textos son largos es preferible no considerar las palabras vacías, no aplicar *stemming* y medir la relevancia de los términos con TF-IDF. En el caso de los textos cortos, no deben eliminarse las palabras vacías, es recomendable la aplicación de *stemming* y se obtienen mejores resultados cuando se mide la importancia de los términos con una medida basada en su frecuencia relativa con respecto al corpus. Las publicaciones en las que se sustentan estas aserciones son (Martínez Cámara et al., 2011; Martínez-Cámara et al., 2015).

La principal aportación en el ámbito de la clasificación no supervisada se encuentra

[1] http://sinai.ujaen.es/cost-2/

[2] http://sinai.ujaen.es/isol/

[3] http://sinai.ujaen.es/coah/

en el diseño de un método modular e independiente del idioma, que está basado en la incorporación al proceso de clasificación de la polaridad de conceptos relacionados con los presentes en el texto. La experimentación desarrollada sobre textos procedentes sobre *tweets* fue bastante positiva (Montejo-Ráez et al., 2014). Mientras que la evaluación que se realizó sobre textos largos demostró que hay que seguir trabajando en encontrar cual es la mejor manera de introducir un mayor grado de información en el proceso de clasificación de la polaridad (Martínez Cámara et al., 2013).

La tesis también ofrece como resultado un método de adaptación al dominio para listas de palabras de opinión. Los experimentaciones que se han realizado para adaptar a iSOL a diversos dominios han demostrado su validez (Molina-González et al., 2014).

En el ámbito de la combinación de clasificadores, las diversas experimentaciones que se han realizado han demostrado que es positiva la combinación de clasificadores especializados en español e inglés, y que el método de combinación más recomendable es el conocido como *stacking* (Martínez-Cámara et al., 2014).

Bibliografía

Martínez-Cámara, E., M. T. Martín-Valdivia, M. D. Molina-González, y J. M. Perea-Ortega. 2014. Integrating Spanish lexical resources by meta-classifiers for polarity classification. *Journal of Information Science*, 40(4):538–554.

Martínez Cámara, E., M. T. Martín Valdivia, M. D. Molina González, y L. A. Ureña López. 2013. Bilingual experiments on an opinion comparable corpus. En *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, páginas 87–93, Atlanta, Georgia, June. ACL.

Martínez Cámara, E., M. T. Martín Valdivia, J. M. Perea Ortega, y L. A. Ureñna López. 2011. Técnicas de clasificación de opiniones aplicadas a un corpus en español. *Procesamiento del Lenguaje Natural*, 47(0):163–170.

Martínez-Cámara, E., M. T. Martín-Valdivia, L. A. Ureña López, y R. Mitkov. 2015. Polarity classification for Spanish tweets using the COST corpus. *Journal of Information Science*, 41(3):263–272.

Molina-González, M. D., E. Martínez-Cámara, M. T. Martín-Valdivia, y L. A. Ure na López. 2014. A Spanish semantic orientation approach to domain adaptation for polarity classification. *Information Processing & Management*, 51(4):520–531.

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.

Molina-González, M. D., E. Martínez-Cámara, M. Martín-Valdivia, y L. Ureña López. 2014. Cross-domain sentiment analysis using Spanish opinionated words. En *Natural Language Processing and Information Systems*, volumen 8455 de *Lecture Notes in Computer Science*. Springer International Publishing, páginas 214–219.

Montejo-Ráez, A., E. Martínez-Cámara, M. T. Martín-Valdivia, y L. A. Ureña-López. 2014. Ranked wordnet graph for sentiment polarity classification in twitter. *Computer Speech & Language*, 28(1):93–107.