 ISBN: 978-84-608-1989-9

Análisis de Sentimientos a nivel de aspecto y estudio de la negación en opiniones escritas en español*

Sentiment Analysis at aspect level and study of negation in Spanish reviews

Salud M. Jiménez Zafra
Departamento de Informática, Escuela Politécnica Superior de Jaén
Universidad de Jaén, E-23071 - Jaén
sjzafra@ujaen.es

Resumen: El análisis de opiniones es una tarea que ha suscitado un gran interés en la comunidad científica en los últimos años, ya que cada día son más las empresas, consumidores, gobiernos, etc. interesados en conocer la opinión que los usuarios tienen acerca de determinados productos, servicios, temas... Pero a esta tarea todavía le quedan muchos frentes abiertos para que se pueda considerar resuelta, como por ejemplo el tratamiento de la negación y el análisis a nivel de aspecto. La mayor parte de la investigación existente sobre estos fenómenos se centra en opiniones escritas en inglés. Por ello, en este trabajo, se describe un proyecto de tesis que se va a centrar en el tratamiento de estos fenómenos en español con el fin de realizar un avance importante en esta área.
Palabras clave: Análisis de opiniones, identificación del ámbito de la negación, análisis a nivel de aspecto.

Abstract: Sentiment Analysis is a task of great interest for the research community in the last years, because every day there are more companies, consumers, governments, etc. interested in the opinion that users have about certain products, services, topics... But there are several issues that have not been sufficiently studied and that some authors consider challenges, such as the treatment of negation and the aspect based sentiment analysis. Most of the research about these phenomena is oriented to documents written in English. Therefore, in this work is described a thesis project that will focus on the treatment of these phenomena in Spanish in order to make a breakthrough in this area.
Keywords: Sentiment analysis, negation scope identification, aspect based sentiment analysis.

1 Introducción

En este trabajo se presenta un proyecto de tesis que tiene como objetivo el estudio de dos de los grandes desafíos del análisis de opiniones: el tratamiento de la negación y el análisis a nivel de aspecto. La mayor parte de las investigaciones realizadas hasta el momento sobre estos fenómenos se centran en opiniones escritas en inglés. Pero hay otros idiomas, entre los que se encuentra el español, que cada día están más presentes en Internet. Un adecuado tratamiento de estos fenómenos supondría un gran avance en esta área. Por ello, esta investigación se va a centrar fundamentalmente en textos en español. El resto del trabajo se organiza como sigue. En primer lugar se mostrarán los motivos que han llevado a la elección de este tema para la realización de esta tesis. A continuación, se realizará una breve revisión de los antecedentes y trabajos relacionados. Posteriormente se describirá la investigación que se va a llevar a cabo y por último se mostrará la metodología a seguir.

2 Motivación

La posibilidad de generar e intercambiar contenido en la web ha suscitado un gran interés por conocer las opiniones que se comparten en este medio. Cada día son más las empresas interesadas en la opinión que los usuarios

* Este trabajo ha sido parcialmente financiado por el Fondo Europeo de Desarrollo Regional (FEDER), el proyecto ATTOS (TIN2012-38536-C03-0) del Gobierno de España, el proyecto AORESCU (TIC-07684) del Gobierno regional de la Junta de Andalucía y el proyecto CEATIC-2013-01 de la Universidad de Jaén.

tienen acerca de sus productos o servicios para determinar qué deben mejorar, qué deben eliminar y qué deben mantener. Además, esta información se ha convertido en un recurso indispensable en la toma de decisiones y en la definición de las estrategias de marketing. Pero este conocimiento no sólo ha originado interés en las empresas sino que los propios consumidores, antes de adquirir un producto o contratar un servicio, utilizan la web para buscar opiniones de otros usuarios. Se trata de una información muy útil que se puede emplear incluso para predecir los resultados de unas elecciones, el éxito de una película... La gran cantidad de fuentes y el elevado volumen de textos con opiniones hacen que resulte complicado para el usuario seleccionar información de su interés. Por ello, es necesario desarrollar sistemas automáticos que faciliten esta tarea, es decir, sistemas que se encarguen de extraer, clasificar y presentar estas opiniones. La disciplina conocida como Minería de Opiniones (MO) o Análisis de Sentimientos (AS) surge para dar solución a este problema. La MO es una disciplina que combina técnicas de Procesamiento del Lenguaje Natural (PLN) y de la Lingüística Computacional para detectar la información subjetiva de un texto y clasificarla. El amplio abanico de aplicaciones en las que se puede emplear ha provocado un gran interés por parte de la comunidad científica, por lo que existen muchos trabajos centrados en este tema, la mayoría de ellos en inglés, pero son muchos los frentes que aún siguen abiertos y que requieren un estudio profundo, como son el tratamiento de la negación, el análisis a nivel de aspecto, el tratamiento de la ironía y del sarcasmo... Algunos autores los definen incluso como desafíos (Pang y Lee, 2008) (Liu, 2012). Un correcto tratamiento de estos fenómenos supondría un avance importante en este área. Por ello, el objetivo de esta tesis es ir un paso más allá de los sistemas tradicionales para tratar de dar solución, en la medida de lo posible, a dos de estos desafíos, el tratamiento de la negación y el análisis a nivel de aspecto. Además, en contraposición de la mayoría de los estudios existentes hasta el momento se va a realizar sobre español ya que su presencia en Internet es cada vez mayor, lo que pone de manifiesto la necesidad de su tratamiento.

3 Antecedentes y trabajos relacionados

3.1 Negación

El tratamiento de la negación es un problema abierto dentro del PLN en general, y dentro de la MO en particular. Se trata de un fenómeno lingüístico que no ha sido estudiado suficientemente y que requiere un análisis profundo. Hasta ahora, la mayor parte de las investigaciones relacionadas con el tratamiento de la negación en el AS se han realizado sobre opiniones escritas en inglés. Las primeras aproximaciones comenzaron en el año 2001 y sugieren métodos relativamente sencillos. Das y Chen (2001) proponen añadir "NOT" ("NOT_word") a las palabras de la oración que se encuentren próximas a términos negativos, como por ejemplo "no" o "don't". Pang, Lee, y Vaithyanathan (2002) siguen un enfoque similar al anterior pero considerando que las palabras afectadas por la negación son todas aquellas que aparecen después del término negativo hasta encontrar el primer signo de puntuación. Estos autores realizan experimentos utilizando algoritmos de aprendizaje automático para comprobar si la clasificación de opiniones teniendo en cuenta la negación mejora, llegando a la conclusión de que con el método propuesto se produce una mejora insignificante. En 2004, a la vista de los resultados obtenidos hasta el momento, Polanyi y Zaenen (2004) dan un paso más allá y tienen en cuenta además de la negación, intensificadores y atenuantes. Además, presentan el primer modelo que asigna puntuaciones a palabras de opinión, invirtiendo la polaridad de las expresiones negadas. Desafortunadamente este modelo no se llegó a implementar por lo que sólo podemos especular sobre su efectividad. Posteriormente, Kennedy e Inkpen (2006) desarrollan un modelo de negación muy similar al propuesto por Polanyi y Zaenen (2004), en el que definen como ámbito de una palabra negativa/intensificador/atenuante aquella inmediatamente posterior. En el caso de las palabras afectadas por la negación siguen un enfoque basado en invertir la polaridad de las mismas, mientras que en el caso de las palabras que se encuentran en el ámbito de intensificadores/atenuantes, lo que hacen es incrementar/disminuir el grado de positividad/negatividad según sea el caso. Para clasificar las opiniones emplean dos métodos, el

primero de ellos consiste en clasificar un comentario en función del número de palabras de opinión positivas y negativas que contiene y el segundo se basa en el uso del algoritmo de aprendizaje automático SVM, llegando a la conclusión de que el tratamiento de la negación es un hecho importante. Por otro lado, Wilson, Wiebe, y Hoffmann (2005) proponen utilizar una ventana fija de tamaño 4 para determinar el ámbito de la negación. Los trabajos presentados son los pioneros en el modelado de la negación en el AS en inglés, pero la comunidad científica sigue trabajando en este tema ya que los enfoques presentados hasta ahora no son lo suficientemente precisos. En los últimos trabajos se plantean métodos basados en la definición de reglas lingüísticas a partir de árboles sintácticos (Jia, Yu, y Meng, 2009) (de Albornoz et al., 2012) , y métodos más complejos como el de Taboada, Voll, y Brooke (2008) en el que se definen diferentes reglas para determinar el ámbito de la negación teniendo en cuenta la categoría gramatical de las palabras adyacentes. Incluso, se pueden encontrar excelentes estudios como el de Wiegand et al. (2010) en el que se realiza una revisión del estado del arte del tratamiento de la negación en el AS en inglés y el estudio de Morante y Sporleder (2012) sobre modalidad y negación en lingüística computacional. Por otra parte, la investigación existente en español sobre este tema es muy limitada. El primer trabajo que conocemos es el de Brooke, Tofiloski, y Taboada (2009) en el que utilizan el mismo enfoque que el empleado en su primera versión en inglés (Taboada, Voll, y Brooke, 2008) pero adaptado al español. Vilares, Alonso, y Gómez-Rodríguez (2013) también han trabajado en este reto demostrando que tener en cuenta la estructura sintáctica del texto para el tratamiento de la negación, de la intensificación y de las oraciones subordinadas mejora con respecto a los sistemas puramente léxicos.

3.2 AS a nivel de aspecto

El otro fenómeno que se pretende abordar en esta tesis es el análisis a nivel de aspecto, también conocido como análisis a nivel de característica. Este análisis se centra en la identificación de los aspectos relacionados con la entidad de estudio (ej. Entidad: hotel. Aspectos: limpieza, personal, localización ...) y en determinar si se ha expresado opinión o no sobre ellos y, en caso afirmativo, señalar si ésta es positiva, negativa o neutra. La mayoría de los sistemas existentes hasta el momento realizan un análisis a nivel de documento (Pang, Lee, y Vaithyanathan, 2002) (Turney, 2002) o a nivel de oración (Yu y Hatzivassiloglou, 2003) (Wilson, Wiebe, y Hoffmann, 2005), es decir, determinan la opinión general del tema, producto, persona... de estudio. Sin embargo, el hecho de que la opinión general de un producto sea positiva no quiere decir que el autor piense que todos los aspectos del producto son positivos, ni el hecho de que sea negativa implica que todo lo relacionado con el producto sea malo. Por ello, los usuarios y compañías no se conforman con conocer la opinión general, sino que buscan un conocimiento más detallado. Al igual que en el caso de la negación, la mayoría de los trabajos existentes sobre este fenómeno se han realizado sobre opiniones escritas en inglés. Los principales métodos que se han empleado para la identificación de aspectos son los basados en la extracción de nombres frecuentes (Hu y Liu, 2004), (Blair-Goldensohn et al., 2008), (Long, Zhang, y Zhut, 2010), en la extracción a partir de palabras de opinión (Hu y Liu, 2004), (Zhuang, Jing, y Zhu, 2006), (Qiu et al., 2011), utilizando métodos de aprendizaje supervisado (Liu, Hu, y Cheng, 2005), (Yu et al., 2011), (Marcheggiani et al., 2014) y empleando modelos de identificación de temas (Mei et al., 2007), (Li et al., 2010), (Sauper, Haghighi, y Barzilay, 2011). Para determinar si la opinión expresada sobre un aspecto es positiva, negativa o neutra la clave se encuentra en determinar correctamente las palabras que se han utilizado en la oración para hablar de ese aspecto. Para ello se han utilizado fundamentalmente métodos basados en analizadores de dependencias (Boiy y Moens, 2009), (Thet, Na, y Khoo, 2010), (Jiang et al., 2011). En los últimos años se está mostrando un especial interés sobre este fenómeno, incluso se ha propuesto como tarea por primera vez en la edición 2014 del workshop SemEval[1] (Pontiki et al., 2014). Pero la investigación sobre este tema en español es prácticamente inexistente aunque empieza a tomar fuerza. En la edición 2014 del taller TASS[2] se ha incluido por primera vez una tarea para el análisis de aspectos en tweets en español (Villena-Román et al., 2015).

[1] http://alt.qcri.org/semeval2014/task4/

[2] http://www.daedalus.es/TASS2014/tass2014.php

4 Descripción de la investigación propuesta

Como se ha mencionado anteriormente, el tratamiento de la negación y el análisis a nivel de aspectos son dos de los grandes retos con los que la comunidad científica se ha encontrado en los últimos años. Estos fenómenos se han empezado a estudiar en la lengua inglesa pero en la lengua española no existe apenas investigación sobre los mismos. Por ello, el objetivo de esta tesis es realizar una propuesta que permita abordar estos fenómenos en español, ya que un correcto tratamiento de los mismos supondría un avance en el AS en particular y en el PLN en general. El punto de partida de esta investigación se encuentra en el análisis de los trabajos existentes hasta el momento en inglés. Este análisis es clave para el inicio de la investigación en español, ya que como una primera aproximación se pretende reproducir algunos de los enfoques más utilizados en inglés con el objetivo de comprobar cómo funcionan en español. Teniendo en cuenta que la negación es un fenómeno lingüístico y que para determinar las palabras que se utilizan al hablar de un determinado aspecto influye la estructura de la oración, un enfoque bastante útil sería aquel que tuviera en cuenta las relaciones sintácticas. Por ello, en esta investigación se van a proponer métodos basados en el análisis de árboles sintácticos tanto para determinar el ámbito de la negación como para identificar las palabras que se utilizan en una oración para hablar sobre un determinado aspecto. Además, debido a la inexistencia de corpus etiquetados a estos niveles en español (negación y aspectos), uno de los objetivos es etiquetar un corpus a nivel de negación para así poder determinar dónde está la fortaleza de cada uno de los sistemas estudiados, es decir, si ésta se encuentra en la identificación del alcance de la negación o en el método de clasificación utilizado. Otro de los objetivos es etiquetar un corpus a nivel de aspecto para identificar las debilidades y fortalezas de los sistemas planteados, con el fin de determinar si se identifican correctamente las palabras utilizadas para hablar sobre un determinado aspecto y si el método de clasificación empleado es adecuado o no.

5 Metodología

La metodología que se propone para la consecución de esta tesis se presenta a continuación:

1. Estudio y revisión del estado del arte. Se comenzará con el estudio y análisis de la bibliografía existente sobre la temática.
2. Adaptación de recursos existentes para poder realizar un análisis de los métodos propuestos.
3. Desarrollo de un prototipo.
   - Diseño de una arquitectura modular que permita integrar nuevas funcionalidades a medida que se vaya avanzando en la investigación.
   - Construcción de la arquitectura modular diseñada.
   - Prueba del correcto funcionamiento del prototipo.
4. Experimentación y evaluación. Se utilizarán los recursos generados para llevar a cabo la experimentación y posteriormente se procederá a la evaluación del prototipo, llevando a cabo una comparación de los resultados obtenidos con los ya existentes. Los resultados obtenidos se pondrán a disposición de la comunidad científica.

Bibliografía

Blair-Goldensohn, Sasha, Kerry Hannan, Ryan McDonald, Tyler Neylon, George A Reis, y Jeff Reynar. 2008. Building a sentiment summarizer for local service reviews. En *WWW Workshop on NLP in the Information Explosion Era*, volumen 14.

Boiy, Erik y Marie-Francine Moens. 2009. A machine learning approach to sentiment analysis in multilingual web texts. *Information retrieval*, 12(5):526–558.

Brooke, Julian, Milan Tofiloski, y Maite Taboada. 2009. Cross-linguistic sentiment analysis: From english to spanish. En *RANLP*, páginas 50–54.

Das, Sanjiv y Mike Chen. 2001. Yahoo! for amazon: Extracting market sentiment from stock message boards. En *Proceedings of the Asia Pacific finance association annual conference (APFA)*, volumen 35, página 43. Bangkok, Thailand.

de Albornoz, Jorge Carrillo, Laura Plaza, Alberto Díaz, y Miguel Ballesteros. 2012. Ucm-i: A rule-based syntactic approach for resolving the scope of negation. En

Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation, páginas 282–287. Association for Computational Linguistics.

Hu, Minqing y Bing Liu. 2004. Mining and summarizing customer reviews. En *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, páginas 168–177. ACM.

Jia, Lifeng, Clement Yu, y Weiyi Meng. 2009. The effect of negation on sentiment analysis and retrieval effectiveness. En *Proceedings of the 18th ACM conference on Information and knowledge management*, páginas 1827–1830. ACM.

Jiang, Long, Mo Yu, Ming Zhou, Xiaohua Liu, y Tiejun Zhao. 2011. Target-dependent twitter sentiment classification. En *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, páginas 151–160. Association for Computational Linguistics.

Kennedy, Alistair y Diana Inkpen. 2006. Sentiment classification of movie reviews using contextual valence shifters. *Computational intelligence*, 22(2):110–125.

Li, Fangtao, Chao Han, Minlie Huang, Xiaoyan Zhu, Ying-Ju Xia, Shu Zhang, y Hao Yu. 2010. Structure-aware review mining and summarization. En *Proceedings of the 23rd International Conference on Computational Linguistics*, páginas 653–661. Association for Computational Linguistics.

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

Liu, Bing, Minqing Hu, y Junsheng Cheng. 2005. Opinion observer: analyzing and comparing opinions on the web. En *Proceedings of the 14th international conference on World Wide Web*, páginas 342–351. ACM.

Long, Chong, Jie Zhang, y Xiaoyan Zhut. 2010. A review selection approach for accurate feature rating estimation. En *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, páginas 766–774. Association for Computational Linguistics.

Marcheggiani, Diego, Oscar Täckström, Andrea Esuli, y Fabrizio Sebastiani. 2014. Hierarchical multi-label conditional random fields for aspect-oriented opinion mining. En *Advances in Information Retrieval*. Springer, páginas 273–285.

Mei, Qiaozhu, Xu Ling, Matthew Wondra, Hang Su, y ChengXiang Zhai. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. En *Proceedings of the 16th international conference on World Wide Web*, páginas 171–180. ACM.

Morante, Roser y Caroline Sporleder. 2012. Modality and negation: An introduction to the special issue. *Computational linguistics*, 38(2):223–260.

Pang, Bo y Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

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-Volume 10*, páginas 79–86. Association for Computational Linguistics.

Polanyi, Livia y Annie Zaenen. 2004. Contextual valence shifters. En *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text*.

Pontiki, Maria, Haris Papageorgiou, Dimitrios Galanis, Ion Androutsopoulos, John Pavlopoulos, y Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. En *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, páginas 27–35.

Qiu, Guang, Bing Liu, Jiajun Bu, y Chun Chen. 2011. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1):9–27.

Sauper, Christina, Aria Haghighi, y Regina Barzilay. 2011. Content models with attitude. En *Proceedings of the*

49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, páginas 350–358. Association for Computational Linguistics.

Taboada, Maite, Kimberly Voll, y Julian Brooke. 2008. Extracting sentiment as a function of discourse structure and topicality. *Simon Fraser Univeristy School of Computing Science Technical Report*.

Thet, Tun Thura, Jin-Cheon Na, y Christopher SG Khoo. 2010. Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of Information Science*, página 0165551510388123.

Turney, Peter D. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. En *Proceedings of the 40th annual meeting on association for computational linguistics*, páginas 417–424. Association for Computational Linguistics.

Vilares, David, Miguel A Alonso, y Carlos Gómez-Rodríguez. 2013. Clasificación de polaridad en textos con opiniones en español mediante análisis sintáctico de dependencias. *Procesamiento del lenguaje natural*, 50:13–20.

Villena-Román, Julio, Eugenio Martínez-Cámara, Janine García-Morera, y Salud M. Jiménez-Zafra. 2015. Tass 2014-the challenge of aspect-based sentiment analysis. *Procesamiento del Lenguaje Natural*, 54:61–68.

Wiegand, Michael, Alexandra Balahur, Benjamin Roth, Dietrich Klakow, y Andrés Montoyo. 2010. A survey on the role of negation in sentiment analysis. En *Proceedings of the workshop on negation and speculation in natural language processing*, páginas 60–68. Association for Computational Linguistics.

Wilson, Theresa, Janyce Wiebe, y Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. En *Proceedings of the conference on human language technology and empirical methods in natural language processing*, páginas 347–354. Association for Computational Linguistics.

Yu, Hong y Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. En *Proceedings of the 2003 conference on Empirical methods in natural language processing*, páginas 129–136. Association for Computational Linguistics.

Yu, Jianxing, Zheng-Jun Zha, Meng Wang, y Tat-Seng Chua. 2011. Aspect ranking: identifying important product aspects from online consumer reviews. En *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, páginas 1496–1505. Association for Computational Linguistics.

Zhuang, Li, Feng Jing, y Xiao-Yan Zhu. 2006. Movie review mining and summarization. En *Proceedings of the 15th ACM international conference on Information and knowledge management*, páginas 43–50. ACM.