Procesamiento de Lenguaje Natural, Revista nº 56, marzo de 2016, pp 99-102 recibido 21-12-2015 revisado 11-01-2016 aceptado 03-02-2016

Análisis del Género Discursivo Aplicado a la Clasificación Automática de la Polaridad en Comentarios sobre Productos*

Gender-Discourse Analysis Applied to Automatic Classification of Polarity in Customer Reviews

John Roberto Rodríguez
Universidad de Barcelona
Gran Via de les Corts Catalanes, 585
roberto.john@ub.edu

Resumen: Tesis doctoral en Lingüística Computacional realizada por John Roberto en la Universidad de Barcelona (UB) bajo la dirección de la Dra. Maria Salamó Llorente (Departamento de Matemática Aplicada y Análisis, UB) y la Dra. Maria Antònia Martí Antonín (Departamento de Lingüística, UB). El acto de defensa de la tesis tuvo lugar el 10 de diciembre de 2015 ante el tribunal formado por los doctores Santiago Alcoba Rueda (Universidad Autónoma de Barcelona), Lourdes Díaz Rodríguez (Universidad Pompeu Fabra) y Mariona Taulé Delor (UB). La calificación obtenida fue Excelente *Cum Laude*.
Palabras clave: Análisis de la polaridad, minería de opiniones, género discursivo

Abstract: Ph.D. Thesis in Computational Linguistics, written by John Roberto at the University of Barcelona (UB), under the supervision of Dra. Maria Salamó Llorente (Department of Applied Mathematics and Analysis, UB) and Dra. Maria Antònia Martí Antonín (Department of Linguistics, UB). The author was examined on December 10^{th}, 2015 by a commitee formed by the doctors Santiago Alcoba Rueda (Autonomous University of Barcelona), Lourdes Díaz Rodríguez (Pompeu Fabra University) and Mariona Taulé Delor (UB). The grade obtained was Excellent *Cum Laude*.
Keywords: Polarity analysis, opinion mining, discursive genre

1 Introducción

Esta tesis trata sobre el análisis de la polaridad en comentarios sobre productos, más exactamente, sobre la clasificación de comentarios como positivos o negativos a partir del uso de información lingüística. En la tesis presento un enfoque al análisis de la polaridad basado en el género discursivo de los comentarios. Según este enfoque, primero se identifican los segmentos que caracterizan el género discursivo de los comentarios y, posteriormente, se evalúa la utilidad que cada tipo de segmento tiene para determinar la polaridad de los comentarios.

La tesis se divide en dos partes. En la primera parte, caracterizo los comentarios como un género mediante el análisis de su estructura discursiva y su registro lingüístico. Sobre la base de ambos análisis postulo que los comentarios se componen de tres tipos principales de segmentos: valorativo, narrativo y descriptivo. En la segunda parte de la tesis, utilizo estos segmentos para calcular la polaridad de los comentarios. La hipótesis de partida es que no todos los segmentos que forman parte del género discursivo de los comentarios contribuyen de la misma manera a expresar la polaridad.

2 Caracterización de los comentarios como un género discursivo

En esta primera parte de la tesis analizo la estructura discursiva y el registro lingüístico de los comentarios. El objetivo de ambos análisis es verificar que los comentarios conforman un género discursivo estable, es decir, que comparten las mismas regularidades estructurales, léxicas y morfosintácticas.

2.1 Análisis de la estructura discursiva de los comentarios

El análisis de la estructura discursiva consiste en identificar las regularidades en el tipo y la

* Esta tesis ha sido financiada por una beca de la Generalitat de Catalunya (2010FI_B 00521).

ISSN 1135-5948

distribución de los diferentes segmentos que componen los comentarios sobre productos. Para ello, efectúo una propuesta de segmentación sustentada en los trabajos existentes sobre la metodología para el análisis del género discursivo, la segmentación automática de textos de opinión y las tipologías textuales. En los experimentos valido mi propuesta de segmentación usando un corpus real de comentarios sobre hoteles extraídos de la web de TripAdvisor[1]. Estos comentarios fueron anotados manualmente para obtener la frecuencia de aparición de los tipos de segmentos propuestos.

Los resultados de este análisis me permitieron concluir que los comentarios sobre productos presentan una estructura discursiva relativamente estable caracterizada por la presencia de tres tipos de segmentos:

- **Narrativo:** relata eventos que acompañan la valoración del producto.
- **Descriptivo:** presenta las características que definen el producto.
- **Valorativo:** expresa la actitud del usuario respecto del producto.

2.2 Análisis del registro lingüístico de los comentarios

El análisis del registro lingüístico consiste en identificar las regularidades léxicas y morfosintácticas que caracterizan los comentarios sobre productos. Este análisis lo llevo a cabo a partir de dos tipos de experimentos: un primer grupo de experimentos están orientados a contrastar el registro lingüístico de los comentarios entre sí (análisis intra-textual) y un segundo grupo, entre comentarios y textos periodísticos (análisis inter-textual).

Al contrastar el registro lingüístico de un conjunto representativo comentarios entre sí –empleando como criterio de clasificación tres clases demográficas: edad, sexo y procedencia de los autores de los comentarios–, no fue posible identificar diferencias léxicas destacables que indiquen que estamos ante diferentes tipos de textos. Por el contrario, al contrastar el registro lingüístico de los comentarios con artículos periodísticos, se constató que existen diferencias léxicas e, incluso, diferencias morfosintácticas significativas que indican que estamos ante dos tipos diferentes de textos.

La conclusión general que se desprende de los dos análisis presentados en esta primera parte de la tesis es que los comentarios sobre productos conforman un género discursivo propio caracterizado por presentar una estructura discursiva estable que puede emplearse para calcular la polaridad general de los comentarios.

Los resultados obtenidos en esta primera parte de la investigación aparecen publicados en: Roberto, Salamó, y Martí (2015a), Roberto, Salamó, y Martí (2013), Roberto, Salamó, y Martí (2012) y Roberto, Martí, y Rosso (2011).

3 Cálculo la polaridad de los comentarios

En esta segunda parte de la tesis analizo la polaridad de los comentarios sobre productos. El objetivo de este análisis es determinar la función que cumplen los segmentos narrativo, descriptivo y valorativo en la expresión de la polaridad. Para ello, (1) clasifico de forma automática cada tipo de segmento y (2) evalúo el rendimiento de cada segmento al aplicarlo para calcular la polaridad general del comentario.

Los experimentos presentados en esta parte de la tesis están destinados a evaluar tres métodos alternativos para identificar de manera automática los segmentos discursivos y a calcular el rendimiento, en términos de precisión, que cada uno de ellos presenta para predecir la polaridad de los comentarios. La selección de estos tres métodos de clasificación obedece a la necesidad de tratar cada tipo de segmento según el propósito comunicativo que lo caracteriza.

3.1 Método 1

El primer método determina la función que cumple el segmento valorativo en la expresión de la polaridad.

Con este fin, selecciono un conjunto de rasgos lingüísticos que utilizo como atributos de entrenamiento para realizar una clasificación supervisada de los tres tipos de segmentos. Estos rasgos describen algunas de las propiedades morfosintácticas y léxicas más características de cada tipo de segmento. Aplicando una aproximación supervisada basada en el uso de bolsa de palabras (BoW) y otra no supervisada basada en la herramienta SO-CAL[2], contrasto el rendimiento que cada tipo

[1] https://www.tripadvisor.es/

[2] SO-CAL es un software para la clasificación no

de segmento presenta al ser usado para calcular la polaridad del comentario completo.

Los experimentos que llevé a cabo en esta parte del análisis me permitieron determinar que es posible aislar de forma automática, con una precisión promedio del 80 %, los segmentos valorativo, narrativo y descriptivo mediante el uso de un conjunto de características léxicas y morfosintácticas. Adicionalmente, he podido comprobar que los segmentos valorativos expresan la polaridad de los comentarios de manera más efectiva que el comentario entero o que los otros segmentos de forma aislada.

3.2 Método 2

El segundo método determina la función que cumple el segmento narrativo en la expresión de la polaridad.

Con este fin, identifico las secuencias narrativas que componen el comentario. Para detectar dichas secuencias narrativas, e inspirado en los trabajos de Chambers (2011), implemento un algoritmo que extrae las oraciones que en un comentario «narran» eventos relacionados temporalmente, es decir, las oraciones que conformarán las secuencias narrativas del texto de opinión. Una vez recuperadas estas secuencias, realizo varios experimentos encaminados a determinar el impacto que su omisión tiene a nivel del cálculo de la polaridad de los comentarios.

Los experimentos que llevé a cabo en esta segunda parte del análisis de la polaridad me permitieron determinar que es posible recuperar de forma automática el segmento narrativo seleccionando las secuencias narrativas que forman parte de los comentarios. Además, he observado que los usuarios recurren a las narraciones para comentar aspectos negativos del producto valorado como un mecanismo transversal de expresión de la polaridad.

3.3 Método 3

El tercer método determina la función que cumple el segmento descriptivo en la expresión de la polaridad.

Con este fin, recupero las oraciones del comentario que describen las características positivas y negativas de un producto (ej. «un airbag de conductor con una forma optimizada para proporcionar una mayor eficacia»), es decir, el segmento descriptivo. Aplicando aprendizaje supervisado, clasifico estas oraciones como simétricas o asimétricas, según expresen o no la misma polaridad que la del comentario. Para el entrenamiento de estos clasificadores utilizo como atributos diferentes índices de la complejidad sintáctica como son los índices de Yngve (Yngve, 1960), Frazier (Frazier y Clifton, 1998) y Pakhomov (Pakhomov et al., 2011). Posteriormente, realizo una serie de experimentos orientados a determinar el impacto que la omisión de las oraciones descriptivas asimétricas tiene sobre la polaridad de los comentarios.

Los experimentos que llevé a cabo en esta última parte del análisis de la polaridad me permitieron constatar que es posible usar la complejidad sintáctica para diferenciar entre oraciones simétricas y oraciones asimétricas. También he observado que la omisión de las oraciones asimétricas (representadas por las oraciones sintácticamente complejas) mejora la detección de la polaridad de los comentarios, especialmente la de los comentarios con polaridad negativa. Este hecho indica que los usuarios se suelen valer de las estructuras sintácticamente complejas para expresar opiniones con una polaridad opuesta a la del comentario.

Los resultados obtenidos en esta segunda parte de la investigación aparecen publicados en: Roberto, Salamó, y Martí (2015a), Roberto, Salamó, y Martí (2015b) y Roberto, Salamó, y Martí (2014).

4 Conclusiones

Las principales conclusiones obtenidas con esta tesis son las siguientes:

- Los comentarios sobre productos poseen una estructura discursiva estable que puede emplearse para calcular su polaridad.
- Los comentarios se componen de tres tipos básicos de segmentos, cada uno de los cuales contribuye de forma diferente y con una intensidad específica en la expresión de la polaridad: valorativo, narrativo y descriptivo.
- El segmento valorativo se usan para expresar la polaridad general del comentario puesto que, como se ha demostrado mediante los experimentos, este tipo de segmento presenta niveles óptimos de

supervisada de textos de opinión que trabaja a partir de léxicos de polaridad.

precisión en el cálculo de la polaridad empleando un número muy reducido de palabras.

- El segmento narrativo se suelen usar para expresar opiniones negativas puesto que al omitirlo de los comentarios con polaridad negativa, los niveles de precisión en el cálculo de la polaridad se reducen de forma significativa.
- El segmento descriptivo que presenta estructuras sintácticamente complejas suele emplearse para expresar opiniones opuestas a las del comentario: al omitir las oraciones asimétricas los niveles de precisión en el cálculo de la polaridad mejoran significativamente.

En general, las diferencias detectadas entre comentarios positivos y negativos son lo suficientemente importantes como para permitirme afirmar que estamos ante dos tipos de subgéneros discursivos: el subgénero de los comentarios positivos y el subgénero de los comentarios negativos. Esta propiedad de los comentarios sobre productos ha de tenerse presente en cualquier estudio sobre el análisis de su polaridad.

5 Herramientas y recursos

La realización de esta tesis ha dado pie a la creación de las siguientes herramientas y recursos:

- Un corpus de comentarios en castellano sobre hoteles que ha sido anotado con diferente información lingüística y metadatos (HOpinion).
- Una Plataforma en Java para el Análisis de Textos de Opinión (AToP) que se ha implementado con el objetivo de facilitar el análisis automático de comentarios sobre productos.
- Un algoritmo para extraer de los comentarios las oraciones que conforman las secuencias narrativas extendiendo el modelo de Chambers y Jurafsky.
- Un léxico específico del dominio de los hoteles que fue creado de forma semiautomática a partir de los más de 18.000 comentarios que integran el corpus HOpinion.

Bibliografía

Chambers, N. 2011. *Inducing Event Schemas and their Participants from Unlabeled Text.* Ph.D. tesis, PhD Dissertation, Stanford University.

Frazier, L. y C. Clifton, 1998. *Reanalysis in Sentence Processing*, capítulo Sentence Reanalysis, and Visibility, páginas 143–176. Dordrecht: Kluwer Academic Publishers, Cambridge, UK.

Pakhomov, S., D. Chacon, M. Wicklund, y J. Gundel. 2011. Computerized assessment of syntactic complexity in alzheimer's disease: a case study of iris murdoch's writing. *Behavior Research Methods*, 43(1):136–144.

Roberto, J., M. A. Martí, y P. Rosso. 2011. Sistemas de recomendación basados en lenguaje natural: Opiniones vs. valoraciones. *Actas IV Jornadas Tratamiento de la Información Multilingüe y Multimodal (TIMM)*, páginas 45–48.

Roberto, J., M. Salamó, y M. A. Martí. 2012. Análisis de la riqueza léxica en el contexto de la clasificación de atributos demográficos latentes. *Procesamiento de Lenguaje Natural*, 1(48):97–104.

Roberto, J., M. Salamó, y M. A. Martí. 2013. Clasificación automática del registro lingüístico en textos del español: un análisis contrastivo. *Linguamática*, 5(1):59–67.

Roberto, J., M. Salamó, y M. A. Martí. 2014. The function of narrative chains in the polarity classification of reviews. *Procesamiento del Lenguaje Natural*, 52:69–76.

Roberto, J., M. Salamó, y M. A. Martí. 2015a. Genre-based stages classification for polarity analysis. En *The 28th Florida Artificial Intelligence Society Conference (FLAIRS), USA*, volumen 1, páginas 1–6.

Roberto, J., M. Salamó, y M. A. Martí. 2015b. Polarity analysis of reviews based on the omission of asymetric sentences. *Procesamiento del Lenguaje Natural*, 54:77–84.

Yngve, V. 1960. A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, 104(5):444–466.