TRANSFERENCIA DE CONOCIMIENTO EN TECNOLOGÍAS DE LA INFORMACIÓN

Modelo para la detección de polaridad orientado a aspecto. Un caso de estudio en el turismo

Model for aspect-oriented polarity detection. A case study in tourism

Alejandro Ramón Hernández[1], María Matilde García Lorenzo[2]

1-Universidad Central de Las Villas, Cuba. aramon@uclv.cu

2- Universidad Central de Las Villas, Cuba. mmgarcia@uclv.edu.cu

Resumen: El uso de internet y de plataformas digitales para emitir opiniones acerca de servicios y productos va en constante crecimiento, constituyendo una fuente de retroalimentación para las empresas. En la industria del turismo, la satisfacción del cliente es primordial, por lo que se trabaja en la mejora de problemáticas que pueden afectar la calidad de sus servicios. El análisis de los comentarios hechos por los usuarios constituye un indicador directo del nivel de la calidad de los servicios, y de cada uno de los aspectos específicos que influyen en la misma. Un proveedor de servicios con la capacidad de analizar la retroalimentación de sus usuarios en un tiempo prudente y resolver las situaciones que provocan insatisfacción, tiene grandes probabilidades de éxito. El problema está en la dificultad de procesar de forma manual la inmensa cantidad de información generada por los usuarios, y la cantidad de posibles aspectos en los cuales se pudiera trabajar. En este trabajo, se presenta una herramienta automática para el procesamiento de las opiniones de los usuarios y la detección de los distintos aspectos sobre los cuales los usuarios hacen revisiones. Se hace un análisis automático de las opiniones emitidas por estos en TripAdvisor sobre varios hoteles Meliá del complejo turístico Cayo Santa María. Se extraen los aspectos más comunes tratados por

los usuarios y las opiniones sobre estos, su evolución en el tiempo usando técnicas de Procesamiento de Lenguaje Natural y se formula una metodología para realizar análisis similares en el futuro.

Abstract: The use of the Internet and digital platforms to express opinions about services and products is constantly growing, constituting a source of feedback for companies. In the tourism industry, customer satisfaction is paramount, so work is being done to improve problems that may affect the quality of its services. The analysis of comments made by users is a direct indicator of the level of service quality, and of each of the specific aspects that influence it. A service provider with the ability to analyze user feedback in a timely manner and resolve situations that cause dissatisfaction has a high probability of success. The problem lies in the difficulty of manually processing the immense amount of information generated by users, and the number of possible aspects that could be worked on. In this paper, an automatic tool is presented for the processing of user opinions and the detection of the different aspects on which users make reviews. An automatic analysis is made of the opinions issued by users on TripAdvisor about several Meliá hotels in the Cayo Santa María tourist complex. The most common aspects treated by users and the opinions about them are extracted, their evolution over time using Natural Language Processing techniques and a methodology is formulated to perform similar analyses in the future.

Palabras Clave: polaridad, opinión, aspecto, Procesamiento de Lenguaje Natural, Análisis de dependencias

Keywords: polarity, opinion, aspect, Natural Language Processing, Dependency Analysis

1. Introducción

El análisis de sentimiento ha cobrado importancia en los últimos años con el aumento de contenido generado por el usuario y el crecimiento en la cantidad de servicios y productos que se ofrecen online. Inicialmente los principales esfuerzos se centraron en el análisis de la polaridad de las opiniones de los usuarios de manera general; sin embargo, este enfoque es incapaz de captar las opiniones que puede tener el usuario respecto a distintos aspectos de la misma entidad (Toh,Zhiqiang & Su, Jian,2016). Por tanto, es conveniente realizar un análisis de sentimiento más profundo que permita captar las diferentes opiniones de los usuarios sobre el mismo producto o servicio, conocido como Análisis Basado en Aspectos.

Con el crecimiento de la cantidad de revisiones, es prácticamente imposible realizar un análisis manual de las mismas para extraer información valiosa que permita apoyar la toma de decisiones tanto por parte de las empresas que brindan el servicio, como para los usuarios. La existencia de un sistema que permita realizar este análisis de manera automática facilita la interacción cliente-proveedor, de manera que reporta beneficios para ambas partes, por otro lado si este sistema permite que las empresas conozcan cuales son las opiniones de los usuarios de cada una de las características de sus productos, les da la oportunidad de estar en constante mejora de las mismas (Quan, C. & Ren, F. 2014); a su vez los usuarios tendrán la posibilidad de encontrar productos n ómás ajustados a sus expectativas en dependencia de las características más importantes para ellos.

El objetivo de este trabajo es presentar un modelo para el análisis de opiniones de usuarios que cumple con las características antes mencionadas. Para dar cumplimiento al objetivo se diseñó un sistema que va desde la recolección y tratamiento de los datos hasta la presentación de los resultados del procesamiento de las opiniones, de forma tal que ayude a la toma de decisiones para empresas y usuarios.

El análisis de las opiniones se realiza a nivel de aspecto, lo que garantiza un nivel de granularidad que permite que se extraiga la mayor cantidad posible de información de

cada opinión. Los aspectos de cada producto son extraídos de las propias opiniones de los usuarios de manera no supervisada, por tanto, solo se tienen en cuenta los aspectos sobre los cuales realmente los usuarios emiten opinión para cada producto, además esto posibilita que el modelo de procesamiento sea aplicado a distintos contextos sin necesidad de modificación. Por otra parte, el procesamiento de los textos de opinión es basado en técnicas de análisis de dependencias textuales y reglas lingüísticas, por tanto, no es necesario un entrenamiento para aplicarlo a otros productos o servicios, aunque nos apoyamos para el análisis en distintos recursos textuales, por lo que el sistema es dependiente del idioma, sin embargo, actualmente los recursos utilizados se encuentran en la mayoría de los idiomas de mayor uso a nivel global.

Para la validación y visualización de los resultados, se tomó como caso de uso las opiniones de los usuarios de TripAdvisor sobre varios hoteles de la Cayería Norte de Villa Clara. Se realiza el análisis de las opiniones desde su recuperación de TripAdvisor y posterior limpieza hasta la presentación de resultados gráficos para la asistencia en la toma de decisiones.

En las siguientes secciones, se describe el modelo general y cada una de las etapas del mismo y finalmente se hace un análisis de los posibles usos por parte de los usuarios de servicios y decisores empresariales.

2. Modelo de procesamiento

En la Figura 1 se muestra la arquitectura general del modelo. Este se conforma por tres módulos principales; el módulo de recuperación y preprocesamiento de las opiniones que recibe como entrada una URL y devuelve como salida una lista de opiniones preprocesadas recopiladas de las URL proporcionada; el módulo de detección de aspectos y modificadores que recibe como entrada las opiniones preprocesadas y devuelve una lista de aspectos detectados, cada uno con una lista de modificadores asociados y un tercer módulo que es el encargado de calcular la polaridad de la opinión

de los usuarios sobre cada uno de los aspectos detectados, basado en los modificadores asociados a los mismos. Un cuarto módulo se acopla a los tres módulos principales para la visualización de los resultados y la creación de tableros de visualización de la presentación de información útil de apoyo en la toma de decisiones. A continuación, se detallan cada uno de los módulos del modelo.

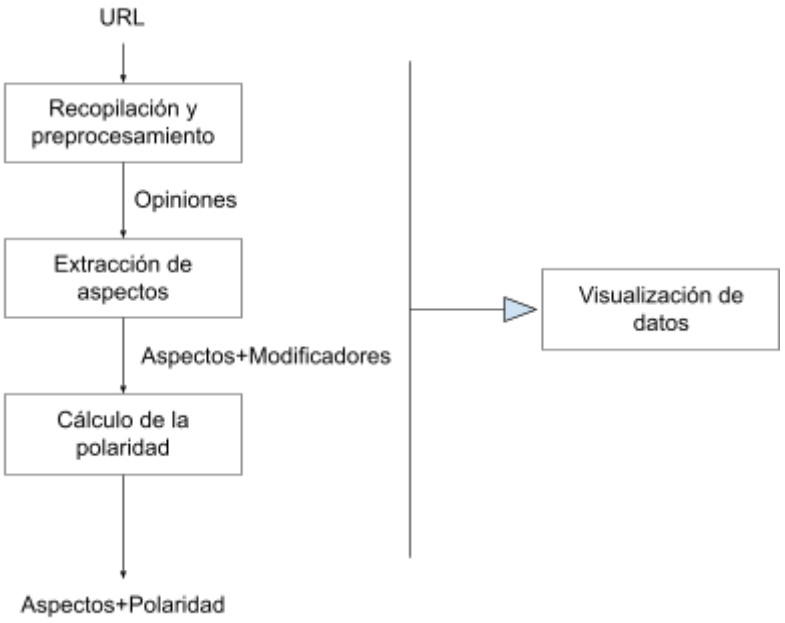

Fig.1 Arquitectura del modelo

2.1. Recopilación y preprocesamiento

Este módulo es el encargado de recuperar las opiniones de la URL proporcionada y aplicar las técnicas de preprocesamiento necesarias para garantizar la calidad de los datos y la efectividad del modelo.

La fase de recopilación de los datos es la única que debe ser modificada para la aplicación del modelo, en dependencia de la URL proporcionada se debe modificar el recolector para que sea capaz de interactuar con la fuente de información, en el caso de

uso que se seleccionó para la validación del modelo se implementó un recolector de datos para extraer la información de TripAdvisor[1] usando la interfaz de Selenium[2] para Python. De esta forma se recuperan las opiniones de los usuarios para cada uno de los hoteles seleccionados para el análisis, de cada opinión se almacena el usuario, la fecha de la estancia en el hotel, la puntuación asignada por el usuario (de 0 a 5) y el texto de la opinión.

La segunda fase de este módulo consiste en el preprocesamiento de las opiniones recuperadas. En el preprocesamiento se modifican los datos recolectados para que sirvan como entrada para las siguientes fases del análisis de sentimiento, una de los modelos de representación de datos más utilizados en la Minería de Texto es el modelo de bolsa de palabras (Radovanovic, M. & Ivanovic, M. 2008), en esta representación los datos innecesarios son eliminados para disminuir su dimensionalidad y complejidad, muchos sistemas utilizan para esto la eliminación de stop words. Otra variante de representación del texto que permite extraer directamente los datos importantes es el etiquetado de los términos en dependencia de su tipo (sustantivo, adjetivo, adverbio, etc), de esta forma se puede hacer un análisis posterior en base a sus dependencias sintácticas (S. Mukherjee & P. Bhattacharyya. 2012). La variante aplicada en este modelo es la de etiquetado de términos y la generación de un árbol de dependencias de cada oración para posteriormente extraer los pares de aspectos-modificadores, para la generación del árbol de dependencia se utilizó *stanza*[3] que es la implementación del equipo de Stanford NLP Group para Python.

Otros pasos de preprocesamiento que se aplican es la detección de n-grams que representan aspectos, para esto se asume que los términos etiquetados como sustantivos que se encuentran consecutivos se refieren a un mismo aspecto de la entidad y la normalización del texto llevándolo todo a minúsculas. La Figura 2 representa la secuencia de pasos de la fase de recolección y preprocesamiento.

[1] https://tripadvisor.com/
[2] https://www.selenium.dev/
[3] https://stanfordnlp.github.io/stanza/

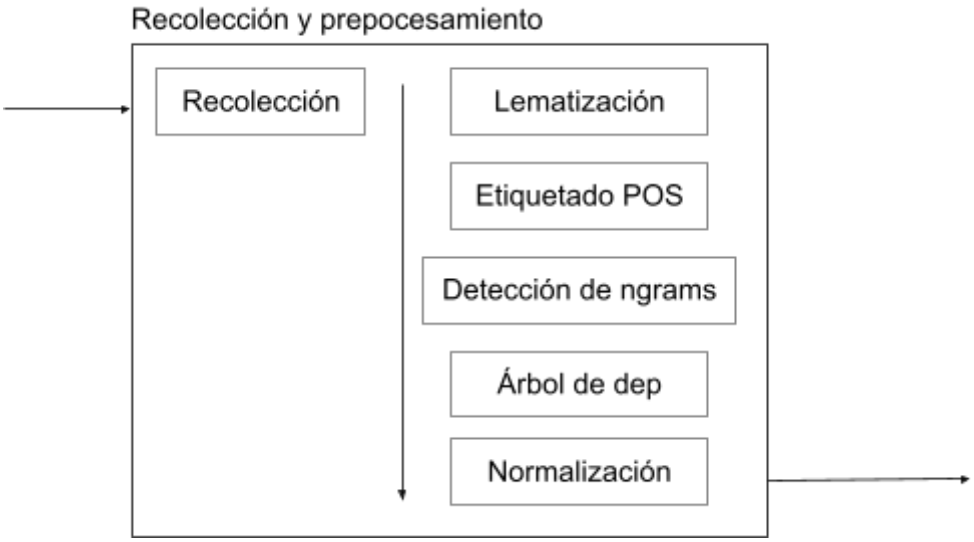

Fig.2 Módulo de recolección y preprocesamiento

2.3. Extracción de aspectos

En esta etapa de procesamiento se toma como entrada las opiniones preprocesadas, primeramente se extraen los aspectos sobre los que potencialmente se emite una opinión para esto se asume inicialmente que todos los sustantivos presentes en la oración pueden considerarse aspectos (Nachiappan Chockalingam. 2018), incluyendo los n-grams previamente detectados como un único término.

Partiendo de los aspectos detectados se identifican en el árbol de dependencias relaciones relevantes que indiquen potenciales modificadores, las relaciones que se tuvieron en cuenta para mantener un balance entre cubrimiento y precisión fueron *nsubj*(relaciones de sujeto nominal, indica el agente en una expresión)[4] y *amod*(relación de adjetivo modificador)[5]. Otro tipo de dependencia que se tuvo en cuenta fue *advmod*(modificador adverbial)[6] en conjunto con una lista de palabras previamente

[4] https://universaldependencies.org/u/dep/nsubj.html
[5] https://universaldependencies.org/u/dep/amod.html
[6] https://universaldependencies.org/u/dep/advmod.html

definidas para la detección de expresiones de negación que afectan el sentido de polaridad de los modificadores. La Figura 3 ilustra los pasos seguidos en esta etapa.

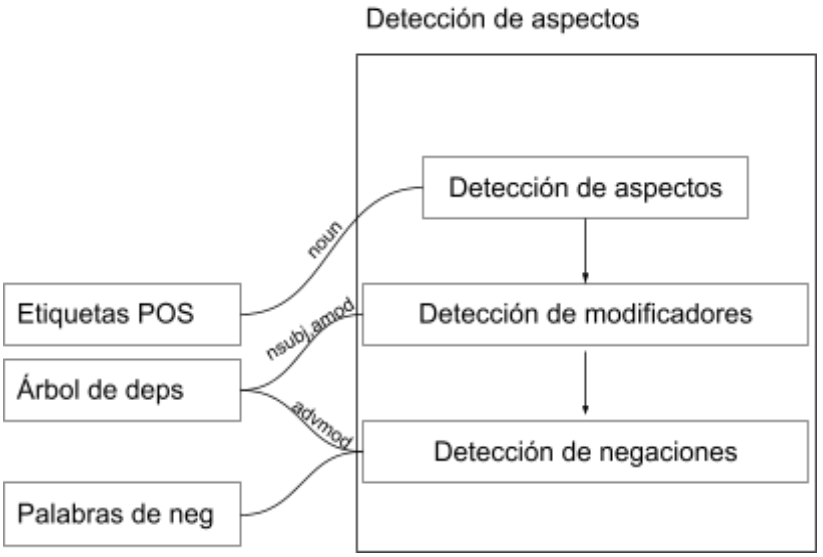

Fig.3 Módulo de extracción de aspectos y modificadores

2.4. Cálculo de la polaridad

El cálculo de la polaridad se realiza para cada uno de los modificadores detectados utilizando el recurso léxico SentiWordNet (Esuli, A. & Sebastiani, F. 2006), que incluye una extensa lista de términos con la polaridad previamente calculada, cada término posee un valor de positividad y otro de negatividad acotados entre cero y uno. A los modificadores que se detectaron en la etapa anterior que están siendo negados se les invierte los valores de polaridad, asignando como polaridad positiva el valor negativo extraído del SentiWordNet y viceversa.

Para determinar la polaridad general de cada aspecto se agregan los valores de las polaridades de cada uno de los modificadores asociados a este. En este modelo, la agregación que se usa es la suma; sin embargo, puede utilizarse cualquier operador de agregación para la combinación de los valores de polaridad.

Los valores de polaridad resultantes son continuos lo que permite que el análisis de las opiniones sobre los aspectos se haga de forma más efectiva conociendo no solo la polaridad de las opiniones de los usuarios sobre el mismo sino también la intensidad de estas.

En la Figura 4 se muestran los pasos seguidos en esta etapa.

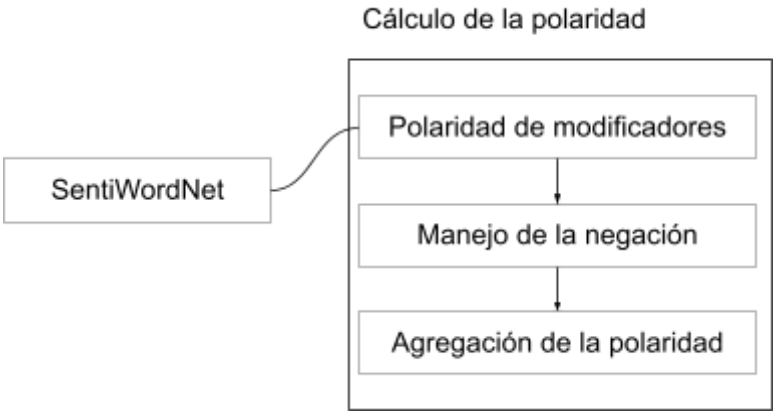

Fig.4 Cálculo de la polaridad

3. Visualización de resultados para el caso de estudio de hoteles Meliá de Cayo Santamaría

Para ejemplificar y visualizar los resultados del modelo se toma como caso de estudio las opiniones de usuarios de TripAdvisor sobre tres hoteles de la compañía Meliá localizados en Cayo Santamaría. Las URLs de donde se extrajo la información son las siguientes:

https://www.tripadvisor.com/Hotel_Review-g670039-d1931078-Reviews-Melia_Buenavista_All_Inclusive_The_Level_Spa-Cayo_Santa_Maria_Villa_Clara_Provi.html
https://www.tripadvisor.com/Hotel_Review-g670039-d13153412-Reviews-Paradisus_Los_Cayos-Cayo_Santa_Maria_Villa_Clara_Province_Cuba.html

https://www.tripadvisor.com/Hotel_Review-g670039-d295233-Reviews-Melia_Cayo_Santa_Maria-Cayo_Santa_Maria_Villa_Clara_Province_Cuba.html

Las opiniones extraídas de la plataforma están en idioma inglés, esto es debido principalmente a que el volumen principal de turismo internacional es de habla inglesa, o al menos se comunica en este idioma, pero representa una muestra considerable de la totalidad de turismo internacional que visita los hoteles estudiados. El siguiente ejemplo muestra una de las opiniones extraídas de la plataforma.

Is a fantastic resort , the staff is very helpful and friendly the Food is good the restaurant are great the pool and beach bar has a great selection the check in is great the information we got help us a lot thru our stay

Los resultados obtenidos del análisis de las opiniones son un indicador de la visión de los clientes hacía distintos aspectos de la empresa, en el caso actual sobre los servicios ofertados por los hoteles. La visualización es esencial para facilitar la interpretación efectiva de los resultados, en el caso de la detección de la polaridad orientada a aspectos resulta de gran utilidad visualizar la puntuación de polaridad de cada uno de los aspectos detectados para identificar puntos fuertes y débiles en el servicio como se presenta en la Figura 5, donde se muestra una relación entre el nivel de polaridad positiva y negativa de cada uno de los aspectos, el tamaño de las barras indica la intensidad de las opiniones, a medida que la barra tiene mayor tamaño quiere decir que la polaridad es más marcada en ese sentido. Se puede apreciar como las opiniones sobre la comida y el servicio son las más positivas mientras la piscina presenta el mayor valor de negatividad aunque de manera general sigue teniendo polaridad positiva, de igual manera pasa con el aspecto comida, esto es debido a que hay gran cantidad de opiniones emitidas sobre estos aspectos, lo que hace que aumenten los valores de positividad y negatividad derivados de las mismas.

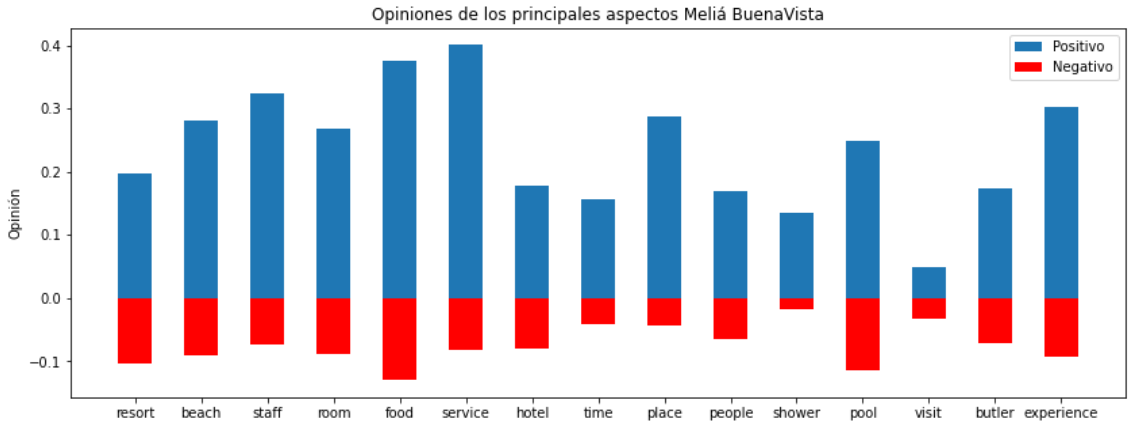

Fig.5 Opiniones de los principales aspectos, hotel Meliá Buena Vista

Otro análisis interesante que se puede realizar es la identificación de los principales términos modificadores asociados a cada aspecto que ayudaría a identificar de forma más detallada los puntos de mayor interés para los usuarios sobre cada uno de los aspectos. En la Figura 6 se muestra una visualización que permite realizar este tipo de análisis, cada término está asociado a un conjunto de modificadores, a medida que el enlace es más fuerte mayor cantidad de usuarios se refieren a esa característica específica del aspecto. Se puede apreciar que para el aspecto *room* los términos modificadores con mayor peso de incidencia son positivos lo que corresponde con el resultado presentado en la Figura 5 donde se muestra cómo para este aspecto la polaridad es mayormente positiva.

Además de estos ejemplos de análisis, muchos otros pueden realizarse tomando como base los resultados del análisis de polaridad, por ejemplo, es posible hacer un seguimiento en el tiempo de la opinión de los usuarios sobre un aspecto específico para medir las diferencias en el estado del mismo y evaluar el impacto de las decisiones tomadas en algún área determinada. Otro posible estudio, es la comparación de los mismos aspectos entre distintas entidades, ya sea para analizar la competencia o, como en el caso del grupo Meliá, analizar el comportamiento de cada una de sus instalaciones bajo los mismos indicadores.

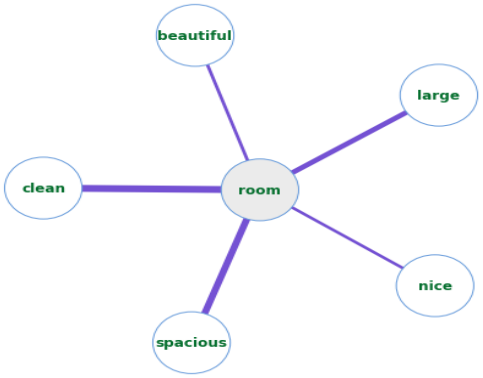

Fig.6 Grafo de relación aspecto-modificadores

Por otra parte, desde el punto de vista de los usuarios el análisis de los distintos aspectos sirve como punto de partida para la selección de un producto o servicio atendiendo a las opiniones de otros usuarios sobre los distintos aspectos de una entidad. En la Figura 7 se muestra un análisis comparativo entre dos hoteles del caso de estudio, atendiendo a aspectos comunes de posible interés para el usuario.

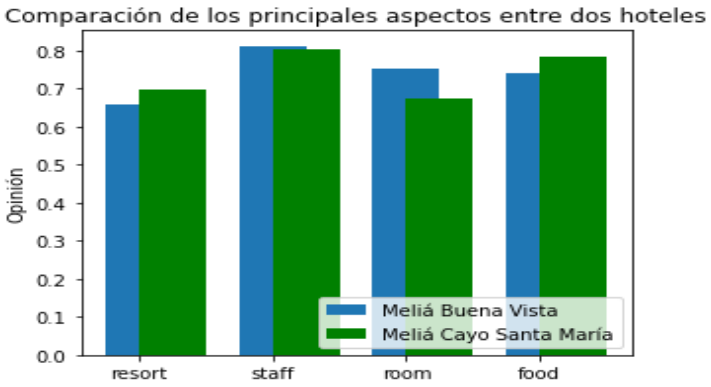

Fig.7 Comparación entre Meliá Santa María y Meliá Buena Vista

4. Conclusiones

El modelo de procesamiento de opiniones desarrollado sirve como herramienta para apoyar la toma de decisiones tanto por parte de los proveedores de servicios y productos como por los usuarios, a partir de los resultados del modelo se puede extraer gran

cantidad de información valiosa en la práctica. Es aplicable a cualquier área de acción donde las opiniones constituyan una fuente de información valiosa, sin necesidad de modificar la estructura general del procesamiento y sin necesidad de disponer de información previa para el análisis, al ser basado en reglas lingüísticas y recursos textuales. Por otra parte, no es capaz de identificar estructuras complejas o mal estructuradas desde un punto de vista lingüístico, aunque se obtienen resultados coherentes y de utilidad práctica y sirve como punto de partida para la realización de procedimientos más avanzados.

Bibliografía

Esuli, Andrea & Sebastiani, Fabrizio. (2006). SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining, *Proceedings of LREC 2006 - 5th Conference on Language Resources and Evaluation*

Nachiappan Chockalingam, Simple and Effective Feature Based Sentiment Analysis on Product Reviews using Domain Specific Sentiment Scores. (2018). POLIBITS, vol. 57, pp. 39–43

Quan, Changqin & Ren, Fuji. (2014). Unsupervised product feature extraction for feature-oriented opinion determination. Information Sciences 272, 16–28

Radovanovic, Milos & Ivanovic, Mirjana. (2008). Text Mining: Approaches and Applications. Novi Sad Journal of Mathematics. 38.

S. Mukherjee & P. Bhattacharyya. (2012). "Feature specific sentiment analysis for product reviews,". International Conference on Intelligent Text Processing and Computational Linguistics. Springer, pp. 475–487.

Zhiqiang Toh & Jian Su. (2016). NLANGP at SemEval-2016 Task 5: Improving Aspect Based SentimentAnalysis using Neural Network Features. Proceedings of SemEval-2016, pages 282–288.