CMA CENTRO DE INVESTIGACIÓN EN METODOS CUANTITATIVOS APLICADOS A LA ECONOMÍA Y LA GESTIÓN IADCOM - UBA

Universidad de Buenos Aires Facultad de Ciencias Económicas

LATINDEX
http://www.latindex.org/latindex/ficha?folio=2751

http://www.economicas.uba.ar/institutos_y_centros/revista-de-investigacion-en-modelos-financieros/

ANÁLISIS EXPLORATORIO DEL EWOM MEDIANTE HERRAMIENTAS DE DATA MINING[1]

Nicolás CHUNG

Departamento de Administración, Facultad de Ciencias Económicas, Universidad de Buenos Aires, Av. Córdoba 2122, 2º Piso – C1120AAQ – CABA – Argentina

nc06031990@gmail.com

Resumen

Recibido: 08/2017

Aceptado: 10/2017

Palabras clave

Datos del consumidor
Métodos estadísticos
Gestión del conocimiento
Datos turísticos

La web 2.0 ha permitido, entre otras cosas, compartir experiencias, opiniones y reseñas acerca de los productos y servicios turísticos que han sido adquiridos. Estas toman la forma de eWOM (comunicación boca-oreja electrónica) y se encuentran en plataformas open data como TripAdvisor y Booking.com, las cuales toman protagonismo hoy en día en el mercado turístico y son un importante factor de decisión de compra.

Este trabajo presenta algunas técnicas provenientes de la minería de datos para procesar una serie de opiniones de TripAdvisor. Con la cual se abordará un análisis exploratorio y se generará información tanto numérica como gráfica, sirviendo como un importante recurso para la toma de decisiones estratégica y/o táctica-operativa.

[1] El título del presente artículo es un extracto de la tesis de maestría del autor, llamada: "Análisis exploratorio eWOM para la gestión de empresas turísticas mediante herramientas de data mining".

ISSN: 2250-687X - ISSN (En línea): 2250-6861

EWOM EXPLORATORY ANALYSIS THROUGH DATA MINING TOOLS

Nicolás CHUNG

Departamento de Administración, Facultad de Ciencias Económicas, Universidad de Buenos Aires, Av. Córdoba 2122, 2º Piso – C1120AAQ – CABA – Argentina

nc06031990@gmail.com

Abstract

KEYWORDS

Customer data
Statistical methods
Knowledge management
Tourism data

The web 2.0 allowed, among many things, to share experiences, opinions and reviews about touristic products and services that have been acquired. These take the shape of eWOM (electronic Word-of-mouth,) and can be found in open-data platforms such as TripAdvisor and Booking.com. Besides, they become significant nowadays in the touristic business and they take part of an important purchase decision factor.

This work presents some data mining techniques to process opinion datasets from TripAdvisor. There, an exploratory analysis will be carried on; and both numeric information and graphics will be generated. Expected to be useful for some decision-making processes, either strategical or operative-tactical.

ISSN: 2250-687X - ISSN (En línea): 2250-6861

1. INTRODUCCIÓN

En el presente trabajo se introducen ciertas técnicas de minería de datos apropiadas para el procesamiento del tipo de datos generado por el usuario en la web 2.0. Este texto generado es conocido como eWOM (comunicación boca-oreja electrónica), y se trata de opiniones, reseñas y experiencias que los usuarios describen acerca de los productos y/o servicios que han consumido.

Al ser estas reseñas difíciles de procesar y gestionar solamente por el factor humano, se debió emplear la ayuda de la tecnología. Por lo tanto, para realizar las tareas de procesamiento de datos, se propuso la utilización de dos softwares: por un lado, el lenguaje de programación y software estadístico R, y por otro, la extensión MeaningCloud de Excel (Microsoft).

Todas las tareas llevadas a cabo consistieron en el procesamiento de datos del tipo texto, el cual presentó ciertos inconvenientes, como errores de tipeo, faltas de ortografía, utilización de lenguaje no estandarizado, giros idiomáticos, jergas, argots, entre otros. Por lo que fueron necesarias sus correcciones para contar con un material estandarizado. Así mismo, también se han unificado ciertos términos cuyos valores semánticos podrían llegar a considerarse como equivalentes; como por ejemplo: el término "habitaciones" se lo modificó por "habitación" (número singular), y el término "cuarto" (muy utilizado en la región rioplatense de América Latina) se lo modificó también por "habitación".

El propósito de estas tareas es el de promover una forma de generación de información que favorezca el proceso de toma de decisiones en una organización. Para ello se realizó el estudio utilizando un solo hotel de gama media y con una cantidad media de opiniones, situado en el sub-barrio Palermo Soho de la Ciudad Autónoma de Buenos Aires (Argentina): Didi Soho Hotel.

Se entiende que: "a mayor información, menor incertidumbre"; y a "menor incertidumbre, mayores las posibilidades triunfar en un negocio". Por lo que, parte de la búsqueda del conocimiento es explotar una fuente de datos Open-data: TripAdvisor. Contando con la ayuda de la tecnología, se espera lograr tener una visión mucho más clara acerca del escenario en el cual las empresas hoteleras y demás empresas turísticas compiten.

2. FORMULACIÓN

Internet es una fuente de información importante para planificar las vacaciones. El 51% de los viajeros lo utilizó en 2010 (iBit, 2011; European Comission, 2010). Los medios sociales, y en general la Web 2.0 permiten a los turistas compartir información en Internet en lo que se llama "leer y escribir la web", en donde el usuario final es al mismo tiempo consumidor y productor de contenidos (iBit, 2011; Nicholas, et al., 2007). Los contenidos generados por los usuarios (User Generated Content - UGC) están teniendo más visibilidad a través de los buscadores (iBit, 2011; Gretzel, 2006), estos quedan depositados en la web y se transmiten en las redes sociales a través del eWOM (boca-oreja electrónico) (iBit, 2011). TripAdvisor crea un espacio en donde se posibilita la interacción C2C (consumidor a consumidor) (Kotler et al., 2011).

Esta nueva pauta de comportamiento del consumidor y las nuevas tecnologías conducen a una mayor transparencia en el mercado (Salvi et al., 2013; Jun et al., 2010). Por lo que, el uso del eWOM es frecuente en el mercado hotelero actual y tiene el potencial para influir en la toma de decisiones de los consumidores (Salvi et al., 2013; Xie et al., 2011). Como resultado, la publicidad boca-oreja electrónica se está sumando a la publicidad boca-oreja tradicional como una influencia de compra importante (Kotler et al., 2011).

Actualmente, los usuarios confían más en las recomendaciones realizadas por otros usuarios que en la propia publicidad (Fernández, 2014; Nielsen, 2013). De hecho, una gran mayoría de los

consumidores cree que estas son últiles, y más de la mitad no reservará en el hotel si no posee opiniones de otros huéspedes (UNWTO, 2014, p. 12) [i]. Por lo que, una sólida y positiva reputación ayudaría a una empresa a lograr una ventaja competitiva y fomentar la repetición de compra (Salvi et al., 2013; Silva y Alwi, 2008). Además, pueden identificar errores en aquellos aspectos que son considerados importantes por los clientes y que dan lugar a la mayoría de sus quejas (Berne et al, 2015; Smyth et al., 2010; Levy et al., 2013). Las empresas turísticas pueden rápidamente tener comentarios positivos o negativos en TripAdvisor, la cual provee una mirada real de las operaciones y la calidad de la gestión de recursos (Fili y Krizaj, 2016).[i]

Cuando se buscan hoteles para alojarse en TripAdvisor, el 80% de los encuestados globales lee entre 6 y 12 opiniones antes de tomar su decisión, y están más interesados en los comentarios recientes que les dan un feedback más actualizado (Hosteltur, 2010). Además, el porcentaje de consumidores que consultan comentarios en TripAdvisor antes de reservar una habitación de hotel ha ido aumentando con el tiempo, así como el número de comentarios que se leen antes de hacer dicha elección (Salvi et al., 2013; Anderson, 2012). Por otro lado, si un hotel aumenta su puntuación en 1 punto en una escala de 5 puntos, el hotel puede aumentar su precio en un 11,2% manteniendo la misma ocupación (Salvi et al., 2013; Anderson, 2012).

Los gestores de marketing del sector turismo y hospitalidad deben entender que sus clientes están expuestos e influenciados por muchos sitios de ventas y opiniones de viajes (iBit, 2011; Litvin, et. al., 2008). Esto permite a los clientes comparar y evaluar estratégicamente los costes y beneficios de las diferentes alternativas, por lo que la oferta en el mercado hotelero es cada vez más compleja (Salvi et al., 2013; Verma, 2010). De esta forma, el alojamiento tiende a ser un producto homogéneo e indiferente y las franquicias son más opacas; desde los años 80 se habla de "comoditización", en donde lo único más importante que el precio del alojamiento, es la ubicación (Hinojosa, 2014; Watkins, 2014). Por tanto, los hoteles pueden aprovechar esta tecnología para mejorar su competitividad (Berne et al, 2015; Buhalis y Law, 2008). Aquellas empresas que no sean capaces de generar valor añadido a través de los sistemas de información se arriesgan a tener que competir por precios, limitando así sus posibilidades de diferenciación (Berne et al, 2015; Olsen y Connolly, 2000).

Por otro lado, el avance tecnológico viene acompañado por un crecimiento desmedido en la cantidad de datos a nivel mundial. La cantidad de datos a nivel mundial generados entre los años 2012 y 2013 superaron la cantidad de datos generados en los años anteriores. Hasta el año 2014 los datos almacenados en total superaron el equivalente a 4,4 millones de millones de giga bytes, y para el año 2020 se prevé que este volumen aumente hasta 10 veces (García, 2014; EMC, 2014). El crecimiento desmesurado de las bases de datos hace necesaria la introducción de nuevas tecnologías junto a la minería de datos (Witter et al., 2011, p.4). En una economía hiper competitiva, centrada en el consumidor y orientada al servicio, los datos son recursos brutos que pueden gatillar el crecimiento de un negocio (Witter et al., 2011, p. 5). [i] La minería de datos sirve para realizar un análisis del mercado efectivo, o bien, comparar los feedbacks de los consumidores para productos similares, descubrir los puntos fuertes y débiles de la competencia, o tomar decisiones de negocios inteligentes… Las tecnologías de inteligencia de negocios (Business intelligence - BI) proveen operaciones de negocio históricas, presentes y con mirada predictiva (Han et al., 2012, p. 27). [i]

Estos datos deben ser interpretados y convertidos en información útil para tomar decisiones razonables (Kotler et al., 2011, p. 137). La comprensión rigurosa de las necesidades del cliente, sus deseos y demandas suministra una información importante para diseñar estrategias de marketing (Kotler et al., 2011, p. 15). Por otro lado, no hay información estratégica más importante que conocer los segmentos que componen el mercado (Levy, 2012, p. 33). Cada segmento demanda un CONES (conjunto esperado de atributos) (Levy, 2012, p. 61). Aprender y

obtener información sirve para reducir la incertidumbre y tomar decisiones más acertadas (Bonatti et al., 2011, p. 70).

De esta forma, futuras investigaciones sobre este tema son necesarias y deben incluir un análisis muy detallado de las opiniones en compañías turísticas (a nivel individual de la compañía y en un segmento específico del turismo) (Fili y Krizaj, 2016).[i]

3. DESARROLLO DEL TRABAJO

El desarrollo del trabajo consiste en cinco partes que muestran técnicas diferentes de la minería de datos para procesar los conjuntos de opiniones. De esta forma, se obtuvieron tanto gráficos como información que resumen los grandes volúmenes de datos.

El conjunto de datos a someter a los experimentos proviene de una instancia de muestreo, que consistió en las 100 primeras opiniones que aparecieran para Didi Soho Hotel en la plataforma. De estas opiniones, se tomaron tanto el título como el texto de la opinión, y también la calificación que el usuario otorga al producto hotelero. Además, se definió que estas deben estar escritas en lengua española.[2]

Estas cinco técnicas se llaman, en orden de aparición: "frecuencia de términos", "asociación de términos", "análisis del sentimiento", "clasificación por subjetividad", y por último, "generación de tópicos". Las dos primeras pertenecen a una rama llamada "minería de textos", mientras que las tres últimas pertenecen a otra rama relacionada llamada "minería de opiniones". Veremos entonces de qué trata cada una de ellas.

3.1. Frecuencia de términos

Witter et al. (2011)[i] explica que la frecuencia de términos consiste en realizar un conteo de todos los términos utilizados en los documentos involucrados. Para ello se utilizó R[3], y el conteo se lo realizó sobre el "texto" y el "título" de las opiniones tomadas.

Para la representación se optó por la utilización del "wordcloud" (traducido como "nube de términos"). Esta popular técnica de visualización de datos imprime cada uno de los términos con las frecuencias más altas y el tamaño de la fuente está directamente relacionado con el nivel de frecuencia.

Una vez realizada la frecuencia de términos, se obtuvo la siguiente representación gráfica con los 50 términos más frecuentes:

[2] Esta muestra se tomó el día 01 de marzo del año 2017.

[3] Paquetes principales utilizados: "NLP", "tm", "wordcloud".

Figura 1. Nube de términos de Didi Soho Hotel.

Los términos más repetidos son: "bien", "ubicación", "habitación", "hotel", y "atención"; estos podrían señalar sobre qué aspectos genéricos del producto hotelero se habló con mayor recurrencia.

Por otro lado, se pueden identificar otros términos con altas frecuencias, los cuales son además adjetivos calificativos; estos son: "cómodo", "grande" y "excelente". De los cuales, por relación semántica, los términos "cómodo" y "grande" estarían más ligados al término "habitación"; así, se podría intuir a priori que:

- "Las habitaciones de Didi Soho Hotel son cómodas y grandes";
- y también, "Las habitaciones podrían ser su ventaja competitiva en el mercado hotelero".

Otras dos posibles relaciones, tomando aquellos términos que más se repiten, son:

- "La ubicación es excelente";
- y, "El personal es amable".

Sin embargo, aun pudiendo contar con esta información, se presentarán a continuación otras técnicas para obtener una visión más nítida acerca de las opiniones, y así, obtener información más certera.

3.2. Asociación de términos

En esta parte, se optó por darle total importancia a la asociación de términos del tipo semántico, que, según lo explicado por Pecina (2009)[i], esta ocurre cuando un par de palabras ocurren en un mismo contexto.

Se realizó entonces la tarea de asociación de términos mediante el lenguaje R[4] para el término "habitación" (el término con mayor frecuencia según la parte 1 – Frecuencia de términos) y se logró confeccionar el siguiente gráfico de nodos (modificado con Adobe Illustrator):

Figura 2. Asociación de términos con el término "habitación"

Visto que la habitación se ve fuertemente relacionada con los términos "grande", "cómodo" y "lindo", y estos a su vez con el término "muy", podría decirse que:

- "las habitaciones son grandes";
- "las habitaciones son cómodas";
- "las habitaciones son lindas";
- "las habitaciones son muy lindas".

Viendo la posible relación entre "grande" y "cómodo", es posible decir que:

- "Las habitaciones son tanto grandes como cómodas".

Con respecto a la relación entre "habitación" y "chico", vemos que a su vez estos dos están relacionados con los términos "grande" y "baño". Por lo tanto, para evitar contradicciones con lo anterior se infiere que:

- "La habitación es grande, pero el baño es chico".

Además, la habitación presenta una fuerte asociación con los términos "desayuno" y "ruido". Ambos podrían señalar lo siguiente:

- "El desayuno se sirve en la habitación";
- y "Desde las habitaciones se pueden escuchar ruidos".

De esta forma, la asociación de términos otorga una rápida visión acerca de un término en particular; esto podría ser utilizado para analizar ciertos aspectos del hotel o del producto hotelero.

[4] Principales paquetes utilizados: "NLP", "tm".

Además, esto podría dar pie a una directa acción por parte del sector operativo, en donde por ejemplo, si los términos "baño" y "sucio" están fuertemente relacionados, tales acciones operativas deberían destinarse de tal forma que los huéspedes ya no perciban el baño como "sucio".

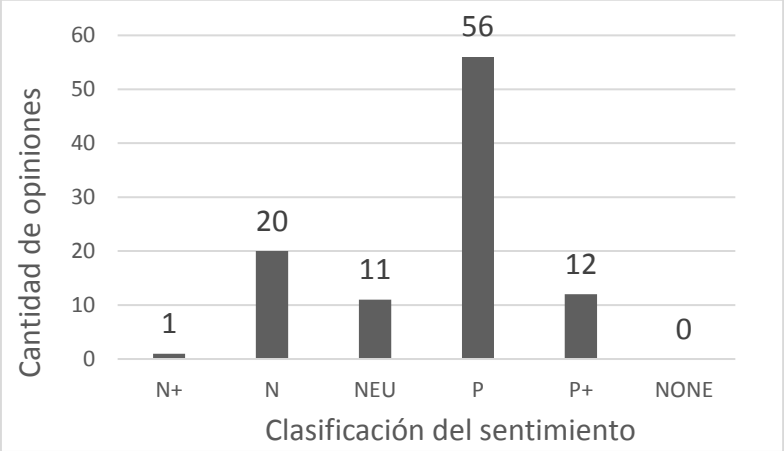
3.3. Clasificación por análisis del sentimiento

Bing (2012, p. 31) explica que la clasificación por sentimiento consiste en clasificar el texto en dos tipos principales de opiniones: opiniones positivas y opiniones negativas, y de acuerdo con las palabras que expresen un sentimiento u opinión. Mientras el vocabulario utilizado tienda a expresar aspectos "positivos" la polaridad tenderá a ser positiva, y si su vocabulario expresa aspectos "negativos" la polaridad tenderá a ser negativa.[i]

Para esta parte, se utilizó la extensión MeaningCloud de Excel, que procesa una celda de una planilla del tipo texto y puede arrojar seis posibilidades, que en orden creciente de polaridad son: N+ (muy negativo), N (negativo), NEU (neutro), P (positivo), P+ (muy positivo), y NONE significa que no es clasificable.

Se sometieron entonces cada una de las opiniones a esta herramienta, y se obtuvieron las siguientes clasificaciones:

Figura 3. Clasificación del texto de la opinión por análisis del sentimiento



Fuente: Elaborción propia.

Como se puede ver, más de la mitad de las opiniones (56 opiniones) fueron catalogadas como "P" (positivo). En segundo lugar, 20 opiniones fueron clasificadas como "N" (negativo), 12 opiniones como "P+", 11 opiniones como "NEU" y 1 sola opinión como "N+".

3.4. Clasificación por análisis de subjetividad

Según Bing (2012), la clasificación por subjetividad se trata de aquella técnica que clasifica un documento, de acuerdo con el vocabulario que utiliza, en "subjetivo" u "objetivo"; en donde lo subjetivo tiene que ver con lo emocional, el juicio de valor, el sentimiento y lo no racional, mientras que lo objetivo tiene que ver con lo racional, el razonamiento, la lógica y lo no emocional.[i]

En esta parte se utilizó add-in "MeaningCloud" de Excel. Esta herramienta arroja tres categorías: "objetivo" (representando un lenguaje racional), "subjetivo" (emocional o irracional), y de no poder clasificar el texto: "sin clasificación".

Previo a someter las opiniones a la herramienta MeaningCloud, se separaron las opiniones por oraciones (tal como se lo hizo en la parte 2 – Asociación de términos). Obteniendo un total de 491 oraciones de un total de 100 opiniones. Teniendo en cuenta que las opiniones pueden estar compuestas tanto de oraciones clasificadas como subjetivas como oraciones clasificadas como objetivas, se procedió a crear un coeficiente de objetividad, en donde, para cada opinión, este coeficiente es el cociente entre la cantidad de oraciones clasificadas como objetivas, y la cantidad total de oraciones de la opinión. Por lo tanto, será expresada mediante la siguiente fórmula:

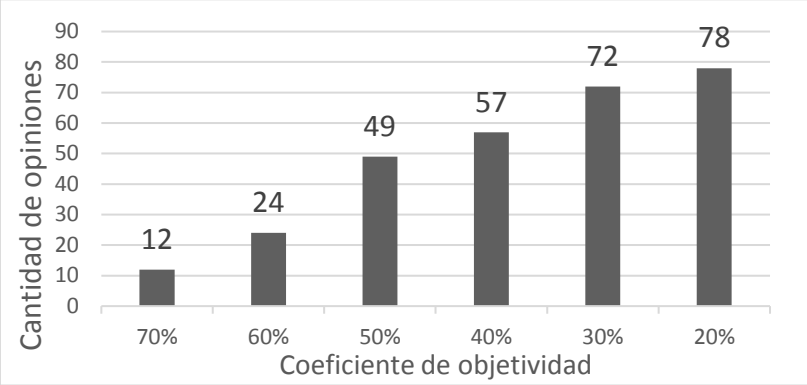
$$Coeficiente\ de\ Objetividad_{[1;100]} = \frac{Q\ oraciones\ objetivas_{[1;100]}}{Q\ total\ oraciones_{[1;100]}}$$

Una vez obtenido este coeficiente para cada una de las oraciones, se procedió a clasificar cada opinión en 6 (seis) variables dicotómicas, en donde los valores pueden ser "0" (cero) para falso y "1" (uno) para verdadero. Estas variables dicotómicas indican si la opinión sobrepasa un determinado porcentaje de objetividad que surge de la totalidad de las oraciones de la opinión, estas van desde el 20% hasta el 70% y van aumentando en 10%[5]. Por lo tanto, la primera variable creada indica que la categoría posee al menos un 20% de oraciones clasificadas como objetivas, la segunda, posee al menos un 30%, y así sucesivamente hasta llegar al 70%.

Una vez que se obtienen estos coeficientes, se consiguieron crear tanto el gráfico, como la tabla pivot en Excel, entre los segmentos y las categorías para el coeficiente de objetividad.

[5] Las categorías de 0%, 10%, 80%, 90% y 100%, fueron descartadas ya que en etapas previas de experimentación, se vio que los valores eran idénticos o muy similares a las categorías adyacentes. Por este motivo se conservó la escala entre el 20% y el 70%.

Figura 4. Clasificación por subjetividad por porcentaje de oraciones clasificadas como objetivas y Opiniones según su coeficiente de objetividad por motivo de viaje



	70%	60%	50%	40%	30%	20%	Total
Amigos	18%	41%	68%	77%	82%	91%	100%
Familia	10%	23%	60%	63%	87%	87%	100%
Negocios	6%	19%	38%	56%	69%	75%	100%
No especifica	17%	17%	17%	17%	33%	50%	100%
Pareja	9%	14%	36%	45%	59%	64%	100%
Solo	25%	25%	25%	25%	50%	75%	100%
Total	12%	24%	49%	57%	72%	78%	100%

Fuente: Elaborción propia.

En la tabla expuesta, se puede decir que el segmento de "amigos" ocupa la cabeza, en donde se pudo ver que un 77% del lenguaje de estos utilizó al menos un 40% de lenguaje objetivo, un 68% del lenguaje utilizó al menos un 50% de lenguaje objetivo, un 41% del lenguaje utilizó al menos un 60% de lenguaje objetivo, y un 18% del lenguaje utilizó al menos un 70% de lenguaje objetivo. También, en menor medida, el segmento "familia" utilizó un 63% del lenguaje con al menos un 40% de lenguaje objetivo para describir su experiencia en el hotel, un 60% del lenguaje utilizó al menos un 50% de lenguaje objetivo, un 23% del lenguaje utilizó al menos un 60% de lenguaje objetivo, y un 18% del lenguaje utilizó al menos un 70% de lenguaje objetivo. Entre los segmentos "negocio" y "pareja", se vio que sus valores son relativamente similares y sus respectivos coeficientes de objetividad son menores a los segmentos de "amigos" y "familia". En cuanto a los segmentos "no especifica" y "solo" no fueron considerados debido a que representan una cantidad de opiniones poco significante.

3.5. Extracción de tópicos

Según Bing (2012, p. 73), la modelación de tópicos asume que cada documento consiste en una mezcla de tópicos o aspectos, y cada uno de estos aspectos puede ser representado por un conjunto de palabras. Así, se identifican los tópicos de colecciones de textos, y simultáneamente, también pueden obtener otro tipo de información acerca de estos.[i]

En otras palabras, se trató de determinar los tópicos o aspectos de cada documento mediante el análisis del texto y el título de la opinión. Este análisis consistió en la identificación de palabras claves que determinen si el usuario se refirió o no a los distintos aspectos. Dado esto, se incluyeron

ocho nuevas variables dicotómicas, las cuales consisten en un análisis del discurso para cada opinión que determina si el usuario se refirió o no a los siguientes aspectos:

- Según TripAdvisor existen seis temas principales: "ubicación", "relación calidad-precio", "calidad de descanso", "limpieza", "personal" y "habitaciones".

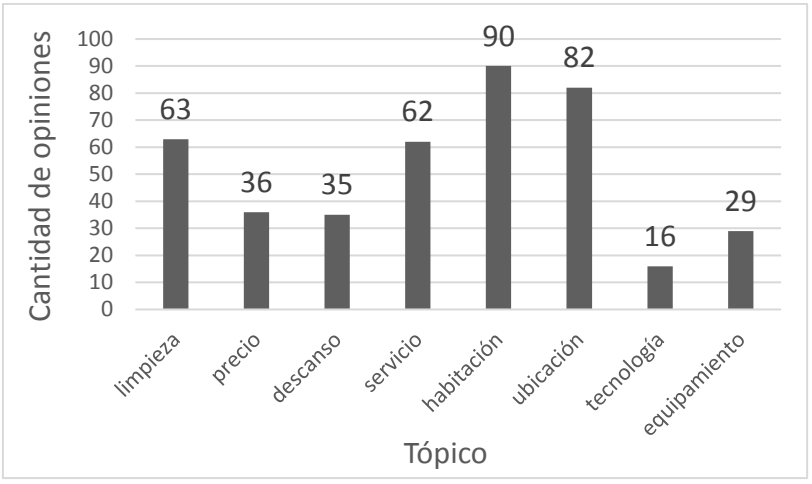
- Además, se agregaron otros dos aspectos para poder contar con más variables:

    - "equipamiento": se trata de todos aquellos objetos extra que posee la habitación; es decir, que aquí se tuvo en cuenta si la habitación posee dispositivos como televisión o aire acondicionado, vajilla o kitchenette, cajones o guardarropas, etc.;
    - "tecnología": señalan tres elementos que relacionan al usuario con la tecnología, estas son: 1) se refirió a la conexión wi-fi; 2) se refirió a la reserva mediante una agencia online; y 3) se refirió a la página TripAdvisor, tanto comentarios anteriores como las fotos que subieron.

Así, mediante la utilización de R[6], se crearon conjuntos de términos para determinar los diferentes aspectos. De esta forma, en todas las nuevas variables dicotómicas, el valor "1" representa que el usuario se refirió al aspecto de la variable correspondiente; mientras que el valor "0", por el otro lado, indica que el usuario omitió este aspecto en su opinión.

En este trabajo, se decidió que un usuario se refirió a este aspecto o tópico, si en el texto de su opinión menciona explícitamente algún término que pertenezca al conjunto de términos de cada tópico. De esta forma, a modo de ejemplo, si el usuario menciona temas relacionados con la "ubicación" del hotel, tales como la zona en la que está situada y/o la cercanía con las atracciones turísticas, y a su vez, utilizando palabras como "zona", "situado" y "cerca", uno de los tópicos (o el único de no ajustarse a más) al cual se ajusta la opinión es la "ubicación", y por lo tanto, se asignó el valor "1" al correspondiente campo de "ubicación" como tópico. Así, entonces, los demás tópicos fueron establecidos.

Figura 5. Cantidad de opiniones por tópicos



Fuente: Elaborción propia.

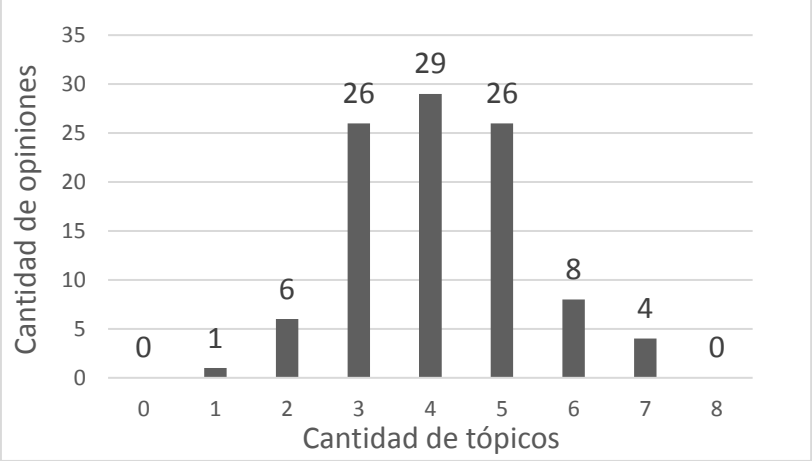
[6] Principales paquetes utilizados: "NLP", "tm".

De esta forma, se obtuvo que los aspectos o tópicos que el usuario más menciona son, en orden descendente: 1) habitación (soportado por el 90% de la muestra); 2) ubicación (soportado por el 82% de la muestra); 3) limpieza (soportado por el 63% de la muestra); 4) servicio (soportado por el 62% de la muestra); 5) calidad de descanso (soportado por el 36% de la muestra); 6) relación precio-calidad (soportado por el 35% de la muestra); y 7) equipamiento (soportado por el 29% de la muestra) ; y 8) tecnología (soportado por el 16% de la muestra). Para el hotel en cuestión:

- "Tanto las 'habitaciones' como la 'ubicación' son los aspectos que más suelen importarle al huésped a la hora de evaluar";
- "Tanto la 'limpieza' como la calidad de 'atención' son aspectos que quedan en segundo plano a la hora de evaluar el producto hotelero".
- "El resto de los aspectos: 'relación precio-calidad', 'calidad de descanso', 'equipamiento' y las facilidades 'tecnológicas' son de menor importancia para el huésped".

Por otro lado, se procedió también a realizar un conteo de aspectos por cada opinión sin discriminar los aspectos, es decir, descubrir cuántos aspectos (sin importar cuáles sean) incluyó un huésped en su opinión. De esta forma, con los mismos aspectos tratados, se obtuvo el siguiente gráfico y utilizando los resultados que arroja R:

Figura 5. Cantidad tópicos no discriminados por opiniones



Fuente: Elaborción propia.

En materia estadística, también calculados con R, se obtuvo que la media fue de 4.13, el desvío estándar: 1.2362, una mediana y una moda de: 4, un coeficiente de asimetría de 0.206 y un coeficiente de curtosis de -0.1252.

Visto cómo se comportan los datos y la información generada, es posible inferir las siguientes proposiciones:

- "Los turistas que se hospedan en hoteles 3 estrellas del barrio de Palermo Soho suelen percibir entre 3 a 5 aspectos del producto turístico";
- "La cantidad de aspectos a la que un turista o huésped suele hacer referencia es una variable que posee una distribución normal".

4. CONCLUSIONES

A lo largo del presente trabajo, se mostró cómo algunas herramientas de la minería de datos nos han ayudado a obtener información cuya apropiada utilización podría llegar a ser útil para modelar un proceso de decisión en una organización. Que debido a la naturaleza del trabajo, las empresas hoteleras serían aquellas organizaciones que más se beneficiarían con estas herramientas.

Las técnicas utilizadas: frecuencia de términos, asociación de términos, clasificación por análisis del sentimiento, clasificación por subjetividad y generación de tópicos, consisten en procesar datos del tipo cadenas de texto y obtener un valor numérico que puede ser representado mediante distintos tipos de gráficos (nube de términos, gráfico de nodos, gráfico de barras).

Estas técnicas son una de las formas de obtener información acerca del contenido generado por el usuario (CGU), el cual puede ser encontrado en las plataformas open data y cuyo contenido podría referirse tanto a las reseñas de los consumidores de un producto o servicio, como también a los otros competidores que ofrecen productos y servicios similares.

Gestionar tanto las herramientas tecnológicas de procesamiento de datos como la información generada a través de estas, no solo significa una reducción del nivel de incertidumbre por obtención de información relevante, sino que resulta ser de suma importancia para contar con un mejor rendimiento de la organización.

5. BIBLIOGRAFÍA

Berne, C., Pedraja, M., Vicuta, A. (2015). El boca-oído online como herramienta para la gestión hotelera. El estado de la cuestión. Vol. 24, 609-626. Universidad de Zaragoza, España. Recuperado de:
http://www.scielo.org.ar/pdf/eypt/v24n3/v24n3a09.pdf

Bing, L. (2012). Sentiment Analysis and Opinion Mining. [Análisis del Sentimiento y Minería de Opiniones]. EUA, Vermont, Williston: Morgan & Claypool Publishers.

Bonatti, P., Aguirre, M., Del Regno, L., Dias, A., Esseiva, F., Lizaso, R., Monti, V., Serrano, S., Slotnisky, A., Tagle, S. y Weissmann, E. (2011). Teoría de la Decisión. Argentina, Ciudad Autónoma de Buenos Aires: PEARSON EDUCACIÓN, S.A.

Feinerer, I. y Hornik, K. (2015). tm: Text Mining Package. R package version 0.6-2. Recuperado de: https://CRAN.R-project.org/package=tm

Fellows, I. (2014). wordcloud: Word Clouds. R package version 2.5. Recuperado de: https://CRAN.R-project.org/package=wordcloud

Fernández, L. (2014). El comportamiento del consumidor online. Factores que aumentan la actividad de búsqueda eWOM en el sector turístico (Tesis de pregrado). Universidad de Oviedo, Oviedo, España.

Fili, M. y Krizaj, D. (2016). Electronic Word of Mouth and Its Credibility in Tourism: The Case of Tripadvisor. [Comunicación boca-oído electrónico y su credibilidad en el Turismo: el

caso de TripAdvisor]. Academica Turistica. Vol. 9(2), 107-111. Recuperado de: http://academica.turistica.si/index.php/AT-TIJ/article/view/64

García, J. (2014, 9 de abril). Los datos en el mundo se multiplicarán por 10 en 2020. La Información. Recuperado de: http://noticias.lainformacion.com/ciencia-y-tecnologia/tecnologia-general/los-datos-en-el-mundo-se-multiplicaran-por-10-en-2020_pGSnrEtEXZYrIhrOcXEy26/

Han, J., Kamber, M., Pei, J. (2012). Data Mining. Concepts and Techniques. [Minería de Datos. Conceptos y Técnicas]. (3ra Ed.). EUA, Massachusetts, Waltham: Morgan Kaufmann Publishers.

Hinojosa, V. (2014, 14 de abril). El riesgo de convertir tu hotel en un commodity. HOSTELTUR. Recuperado de: http://www.hosteltur.com/147890_riesgo-convertir-tu-hotel-commodity.html

Hornik, K. (2016). NLP: Natural Language Processing Infrastructure. R package version 0.1-9. Recuperado de: https://CRAN.R-project.org/package=NLP

HOSTELSUR. (2014, 12 de febrero). TripAdvisor: ¿son fiables los comentarios de los viajeros? Recuperado de: https://www.hosteltur.com/136940_tripadvisor-son-fiables-comentarios-viajeros.html

iBit (2011). Guía metodológica para la gestión de la visibilidad y reputación online de un destino turístico. Caso práctico sobre el destino turístico Calvià (Mallorca). USA, California, San Francisco: Creative Commons. Recuperado de: http://invattur.aimplas.es/ficheros/noticias/1271451453192_ca.pdf

Kotler, P., García de M, J., Flores, J., Bowen, J. y Makens, J. (2011). Marketing Turístico. (5ta Ed.). España, Madrid: PEARSON EDUCACIÓN, S.A.

Levy, A. (2012). Mayonesa. Estrategia, cognición y poder competitivo. (3ra Ed.). Argentina, Ciudad de Buenos Aires: Ediciones GRANICA S.A.

MeaningCloud LLC (2016). MeaningCloud Add-in for Excel. Version 3.1.1.1. Recuperado de: https://www.meaningcloud.com/developer/excel-addin

Microsoft Corporation. (2016). Excel (Student Version 2016). Recuperado de: https://portal.office.com/OLS/MySoftware.aspx

Pang, B. y Lee, L. (2008). Opinion Mining and Sentiment Analysis. [Minería de Opiniones y Análisis del Sentimiento]. EUA, New York, Ithaca, Universidad de Cornell, Departamento de Ciencias de la Computación: NOW Publishers.

Pecina, P. (2009). Lexical Association Measures. [Medidas de Asociación Léxica]. República Checa, Praga, Univerzita Karlova: Instituto de Lingüística Formal y Aplicada.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Recuperado de: https://www.R-project.org/

Salvi, F., Serra, A., Ramón, J. (2013). Los impactos del eWOM en hoteles. REDMARKA. Universidad de las Islas Baleares, España. Recuperado de: http://www.cienciared.com.ar/ra/usr/39/1472/redmarka_n10_v2pp3_17.pdf

Witter, I., Frank, E., Hall, M. (2011). Data Mining. [Minería de Datos]. (3ra Ed.). EUA, Massachusetts, Burlington: Morgan Kaufmann Publishers.