Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Estudios de Posgrado

Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Estudios de Posgrado

MAESTRÍA EN ECONOMÍA Y GESTIÓN DEL TURISMO

PROYECTO
TRABAJO FINAL DE MAESTRÍA

Análisis exploratorio eWOM para la gestión de empresas turísticas mediante herramientas de data mining.
Casos de aplicación: Didi Soho Hotel y Blue Soho Hotel.

AUTOR: NICOLÁS CHUNG

DIRECTOR: JUAN PABLO BALDOMAR

CODIRECTOR: ROBERTO ABALDE

AGOSTO 2017

Resumen (ES)

El presente trabajo realiza un análisis exploratorio de los datos cuantitativos y cualitativos de la plataforma TripAdvisor, estos corresponden a las evaluaciones de consumidores y toman la forma de eWOM (comunicación boca-oído electrónico). La finalidad principal es obtener y/o generar información para la toma de decisiones en un contexto empresarial y turístico.

La investigación es exploratoria, ya que intenta conocer el comportamiento de esta particular fuente de datos y recurso digital; y es del tipo mixta, ya que utiliza ambas metodologías cuantitativa y cualitativa, debido a la naturaleza de los datos. Las tareas de análisis exploratorio fueron llevadas a cabo mediante herramientas informáticas de procesamiento de datos con técnicas de la minería de datos (principalmente de PLN – procesamiento del lenguaje natural).

Entre los resultados obtenidos, se logró exponer a priori: a) el comportamiento estadístico de algunas variables numéricas; b) la relación de los tipos de viajeros con otras variables; c) atributos de los textos escritos por los usuarios; y d) aspectos y apreciaciones de los huéspedes con respecto al producto hotelero.

Mediante el análisis de estos resultados, se logró proponer a priori: a) información para la gestión de los recursos hoteleros; b) información genérica acerca del mercado hotelero; y c) información para la toma de decisiones estratégicas y de competitividad.

Dentro de las limitaciones, se estableció que toda clase de dato recabado, información e hipótesis propuesta corresponde únicamente a los hoteles independientes de gama media situados geográficamente en el sub-barrio de Palermo Soho. En cuanto a la información e hipótesis propuestas simplemente son observaciones a priori de patrones de comportamiento de los datos; de ninguna forma implican una explicación, correlación o causalidad.

Por último, se espera que este tipo de investigación pueda contribuir con la comunidad académica, investigativa y científica de las ciencias económicas y la disciplina turística, contribuir con el mundo empresarial y profesional del campo turístico, y abrir nuevas líneas de investigación dentro y fuera de las disciplinas involucradas.

Palabras claves: eWOM, análisis exploratorio de datos, PLN, gestión de la información.

Abstract (EN)

The current research carries out an exploratory data analysis on quantitative and qualitative data from TripAdvisor.com, gathering consumers' evaluations, acquainted as eWOM (electronic word of mouth). Basically, it intends to both obtain and generate information to support decision-making processes towards businesses and touristic environment.

It is made to be an exploratory research, given that it attempts to discover this particular source of digital data behavior; Also, it is made to be mixed, since it employs both quantitative and qualitative methodologies, because of the data characteristics. Tasks were implemented through some specific data processing software and applying different data mining techniques (mainly: NLP – natural language processing).

Among the obtained results, the following a priori outcomes could be exposed: a) some numeric variables statistic behavior; b) relationships between traveler's type and other variables; c) by-users written text attributes; and d) hotel product aspects and appreciations according to guests' experiences.

Through analyzing the outcomes, the following a priori propositions could be exposed: a) hotel assets management information; b) generic information about the hotel market; and c) strategic and competitive decision-making information.

Regarding the inescapable constraints, each collected datum, and the information and hypotheses proposed may uniquely correspond to the mid-range independent hotels geographically situated in the sub-neighborhood of Palermo Soho. In addition, the proposed information and hypotheses, they are meant to be a priori observations about data behavior; in no way any explanation, correlation or causality will be suggested or implied.

Finally, it is expected that this particular research could contribute to the academic, research and scientific community from the economic sciences and tourism studies as well as make a contribution to the business and professional touristic field; and trigger new studies inside and outside the disciplines involved.

Key words: eWOM, exploratory data analysis, NLP, information management.

Sumário (PT)

Este trabalho faz uma análise exploratória de dados quantitativos e qualitativos da plataforma TripAdvisor, estas correspondem às avaliações dos consumidores e tomam a forma de eWOM (comunicação boca-ouvido eletrónica). O objetivo principal é obter e/ou gerar informação para a tomada de decisões em um contexto de negocios e de turismo.

A pesquisa é exploratória, porque tenta compreender o comportamento desta particular fonte de dados e recurso digital; e é de tipo misto porque usa ambas metodologias quantitativa e qualitativa, devido à natureza dos dados. As tarefas de análise exploratórias foram realizadas por programas de computador de processamento de dados e técnicas de mineração de dados (principalmente PLN - processamento de linguagem natural).

Entre os resultados obtidos, foi possível expor a priori: a) o comportamento estatístico de algumas variáveis numéricas; b) a relacionamento entre os tipos de viajantes com outras variáveis; c) os atributos de textos escritos pelos usuários; e d) os aspectos e as apreciações dos hóspedes sobre o produto do hotel.

Ao analisar esses resultados, foi possível expor a priori: a) informação para a gestão de recursos hoteleiros; b) informação genérica sobre o mercado dos hoteles; e c) informação para a tomada de decisões estratégicas e de competitividade.

Dentro das limitações estabelecido que todos os tipos de dados coletados, informações e hipóteses proposto exclusivamente para os hotéis independentes de médio porte localizados geograficamente no sub-distrito de Palermo Soho. Quanto à informação e suposições são simplesmente propostas de observações a priori de padrões de comportamento de dados; não implica uma explicação ou correlação causal.

Finalmente, espera-se que tais pesquisas possam contribuir para a comunidades académicas, investigativas e científicas das ciências económicas e da disciplina turística, contribuir para o mundo dos negócios e no campo do turismo profissional e abrir novas linhas de pesquisa dentro e fora as disciplinas envolvidas.

Palavras-chave: eWOM, análise exploratória de dados, PLN, gestão da informação.

"El ignorante afirma; el sabio duda y reflexiona"

- *Aristóteles.*

"El conocimiento es poder"

- *Francis Bacon.*

"Nuestra mayor fuente de conocimiento son nuestros clientes más insatisfechos"

"La información es poder"

- *Bill Gates.*

"El verdadero conocimiento es saber la magnitud de la propia ignorancia"

- *Confucio.*

"La lógica te llevará de A a B; la imaginación te llevará a donde sea"

"Lo importante es no dejar de cuestionar; la curiosidad tiene su propia razón de existir"

- *Albert Einstein.*

"Nuestro conocimiento es necesariamente finito, mientras que nuestra ignorancia es necesariamente infinita"

"La verdadera ignorancia no es la ausencia de conocimientos, sino el rehusarse a adquirirlos"

- *Karl Popper.*

"En cualquier momento de decisión lo mejor es hacer lo correcto, luego lo incorrecto, y lo peor es no hacer nada"

- *Theodore Roosevelt.*

"Solo sé que no se nada"

"La verdadera sabiduría está en reconocer la propia ignorancia"

- *Sócrates.*

"Conoce a tu enemigo y conócete a ti mismo; en mil batallas nunca saldrás derrotado"

- *Sun Tzu.*

Diccionario de acrónimos, términos y sinónimos

Angloparlante	Persona, sujeto o usuario que posee habilidades competentes para comunicarse en lengua inglesa
AT	Asociación de términos
α	Probabilidad de cometer el error del tipo I
BI	Business Intelligence (inteligencia de negocios)
CAS	Clasificación por análisis del sentimiento
CGU	Contenido generado por el usuario
CGV	Contenido generado por el viajero
Consumidor	Este término será equivalente a otros como: cliente, usuario, demanda, demandante, internauta, huésped, turista, viajero.
CPS	Clasificación por subjetividad
Empresa	Este término será equivalente a otros como: hotel, hotel-empresa, empresa turística, empresa hotelera, oferta, oferente, negocio.
eWOM	Comunicación electrónica boca-oído (electronic Word of Mouth)
σ	Desvío estándar de la muestra
ς	Desvío estándar de la población
FT	Frecuencia de términos
GRO	Gestión de la reputación online
Hispanoparlante	Persona, sujeto o usuario que posee habilidades competentes para comunicarse en lengua española o castellana
LN	Lenguaje Natural
Lusoparlante	Persona, sujeto o usuario que posee habilidades competentes para comunicarse en lengua portuguesa
MDD	Minería de datos
MDO	Minería de opiniones
MDT	Minería de textos

Me(x)	Mediana
MTC	Medida(s) de tendencia central
Mo(x)	Moda / modo
μ	Media aritmética de la población
$\bar{x}$	Media aritmética de la muestra
OTA	Online Travel Agency
PLN	Procesamiento del Lenguaje Natural
PHNP	Pruebas de hipótesis no paramétricas
PHP	Pruebas de hipótesis paramétricas
R	Lenguaje de programación y software estadístico R
RA	Regla(s) de asociación
RS	Red(es) social(es)
ρ	Índice de correlación de Spearman
.csv	Archivo valores separados por comas

Índice de contenidos

Índice de figuras

INTRODUCCIÓN

1.1. Presentación

El presente trabajo aborda un análisis exploratorio sobre datos recopilados mediante una serie de programas informáticos de procesamiento de datos. Estos datos secundarios provienen de las evaluaciones que los usuarios dejan en TripAdvisor, los cuales describen sus experiencias turísticas (en forma de eWOM – comunicación electrónica boca-oído). Mediante criterios de descarte y reducción del universo, se optó por estudiar dos hoteles específicos: Didi Soho Hotel, como el hotel principal, y Blue Soho Hotel como el hotel secundario, de referencia o competencia, ambos hoteles situados en el sub-barrio de Palermo Soho (barrio de Palermo, Ciudad Autónoma de Buenos Aires, Argentina). Las muestras constan de los 100 comentarios más actuales, los cuales van a partir de finales del año 2009 hasta principios del año 2017.

Este análisis y procesamiento de datos se realizará mediante distintas técnicas estadísticas y de data mining, a fin de exponer el comportamiento de los datos y representar la información generada mediante gráficos mediante el lenguaje R y Excel de Microsoft. Estas técnicas están separadas por secciones dentro de la parte principal de desarrollo, y se tratan de: 1) estadística descriptiva, 2) frecuencia de términos, 3) asociación de términos, 4) análisis del sentimiento, 5) clasificación por subjetividad, 6) generación de tópicos, 7) reglas de asociación, y por último, 8) estadística inferencial.

El propósito por el cual trasciende este trabajo es la intención de mostrar cómo algunas de las técnicas de la minería de datos pueden contribuir a la necesidad de obtención de información por parte de las empresas turísticas; en especial, aquellas relacionadas con el PLN, las cuales procesan el texto generado por el usuario o viajero. De esta forma, al conocer más acerca del contexto en donde estas actúan, los órganos decisores estarán en mejores condiciones de tomar decisiones más acertadas de gestión de recursos y/o de competitividad, o bien, poder afrontar situaciones diversas en el ámbito empresarial.

Como complemento al análisis de los resultados de los datos procesados, se estará proponiendo diversas formas de utilizar la información generada, y también, formas de mejorar la precisión de esta información, o bien, corroborar la que ya fue generada. Además, vale aclarar que otro objetivo no menor que permite la naturaleza exploratoria del trabajo es proponer nuevas líneas de investigación para estudios futuros, dentro o fuera de las disciplinas involucradas.

1.2. Descripción

Primero, la presentación de este trabajo es un requisito de los maestrandos de UBA FCE para obtener el título de Magíster, previamente habiendo obtenido la aprobación del proyecto de este trabajo por parte de la Escuela de Estudios de Posgrado, también de UBA FCE. Esta deberá obedecer los parámetros propuestos por el Instructivo y Protocolo para la presentación del Trabajo Final de Maestría, también dispuesto por la Escuela de Estudios de Posgrado. Segundo, la motivación del maestrando se centra especialmente en realizar un trabajo de investigación que contribuya al conocimiento y la comunidad científica-académica dentro de las ciencias económicas aplicadas en el rubro turístico, aunque también podría suceder en otros rubros.

El tema del trabajo integra diversos campos disciplinarios, siendo las ciencias económicas las predominantes, con fuertes conexiones con la gestión de empresas y el marketing, y el estudio del comportamiento del turista como ejes centrales; y luego la investigación se apoya en otros campos como lo son la estadística, las ciencias de la computación, la lingüística, y la filosofía empresarial, para agregar consistencia a las justificaciones y procedimientos de un trabajo exploratorio con metodología mixta.

La mayoría de las herramientas constan de procedimientos y técnicas de una de las ramas de las ciencias de la computación: la minería de datos (o Data mining en inglés). Dentro de esta también existen otras ramas como la "minería de textos" y la "minería de opiniones", las cuales no han sido ampliamente tratados hasta en la actualidad. Sobre todo, teniendo en cuenta que los datos a procesar no son del tipo numéricos, es decir, se procesa lo que está escrito por una persona, el cual debe a su vez basarse en una lengua clasificada como "natural", que en este trabajo fue el español o castellano.

En este trabajo, estos datos se refieren específicamente a las opiniones que los usuarios dejan en la plataforma de opiniones turísticas TripAdvisor. El cual, por una cuestión de conveniencia, previsión, prevención y falta de trabajos previos, se estableció que las muestras deben ser las 100 más actuales, y estas fueron tomadas el día 1º de marzo del año 2017, conformando un archivo de conjunto de datos de fácil y rápido acceso. Las herramientas utilizadas para procesar los datos fueron principalmente el lenguaje de programación y software estadístico llamado "R", y Excel (Microsoft) junto a un add-in llamado MeaningCloud. De esta forma, se establecieron los recursos para proceder con el procesamiento datos, exposición de resultados a priori, y finalmente, el análisis de los datos por parte del investigador.

1.3. Relevancia

El presente trabajo se inspiró a partir de diversas asignaturas y temas vistos durante la cursada de la maestría correspondiente, con la particularidad que se investigó más acerca de estos para conformar así la base teórica a la cual se estará apoyando el trabajo. Es así como entonces, aquellas asignaturas que más incumben a las ciencias económicas (gestión de empresas/organizaciones y marketing, estadística), y particularmente al comportamiento del turista (tecnología, psicología, antropología y sociología) visto desde distintos ángulos, fueron considerados para el presente trabajo.

De esta forma, se optó por explorar un fenómeno específico que se da en la web: el eWOM, el cual es una forma de comunicación C2C (consumidor a consumidor). Ya que, esta es una fuente de información pública que ayuda a la toma de decisiones tanto de un consumidor para elegir un destino u hotel, como de una empresa para mejorar su imagen en línea o posición en el mercado. Su estudio y la revelación del conocimiento subyacente en este recurso podría llegar a generar un cambio en el mundo turístico-hotelero, en el ambiente económico-empresarial, o bien, en el sector académico y científico.

1.4. Justificación

Para determinar el tema, se tuvo en cuenta los conocimientos previos del maestrando, como lo son las herramientas de procesamiento y análisis de datos (numéricos y no numéricos), la experiencia y la formación en materia de gestión empresarial y la tecnología, y otras no menores como la estadística y la lingüística. De esta forma, se consideró oportuno iniciar el estudio de las opiniones de los turistas, visto especialmente desde la perspectiva gerencial con fines estratégicos y financieros, y bajado a un caso práctico ya que se desconocen trabajos previos que hayan conseguido realizar un abordaje práctico, replicable e integral con diversas herramientas de procesamiento de datos seguido de sus implicancias y aportes en la comunidad científica o empresarial.

Por otro lado, hablando personalmente desde las motivaciones del investigador, se pensó el tema como una forma de introducir y/o iniciar una vocación científica-académica, en donde se proyectó la continuidad con estudios de alto nivel académico, participar en proyectos de investigación multidisciplinarios, y delimitar la carrera profesional dentro de la consultoría o los servicios profesionales.

1.5. Estructura

En cuanto a la estructura del presente trabajo, la secuencia en la cual se presentaron los capítulos del desarrollo no representó ninguna particularidad con respecto a lo que dictaría un modelo estándar de tesis académica de maestría. De hecho, se siguió el formato dispuesto según la Escuela de Estudios de Posgrado de la Facultad de Ciencias Económicas de la Universidad de Buenos Aires.

Por lo que, tendremos así una serie de 8 (ocho) capítulos dentro de la parte de desarrollo del trabajo que representan distintas formas de explorar los datos recolectados, cada uno, además de la aplicación de técnicas y herramientas de procesamiento, con sus respectivas conclusiones, posibles aplicaciones y líneas de investigación futuras. Estos ocho capítulos son, según su orden de presentación: 1) Estadística descriptiva; 2) Frecuencia de términos; 3) Asociación de términos; 4) Clasificación por análisis del sentimiento; 5) Clasificación por subjetividad; 6) Generación de tópicos; 7) Reglas de asociación; 8) Estadística inferencial.

A esto le sigue una parte no menor en este trabajo, que es la parte de conclusiones que complementa al desarrollo, en donde se expone, a grandes rasgos, formas de perfeccionar la información creada, interpretar y utilizar esta información, corroborar la veracidad de la información, sugerir la intervención de otras técnicas, e incluso, otras disciplinas.

PLANTEAMIENTO DEL TEMA

2.1. Formulación del tema

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 electrónico 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 de 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 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 Silva, 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]

La pregunta problematizante será: ¿Cómo la(s) empresa(s) turística(s) puede(n) aprovechar los datos de la plataforma TripAdvisor a su favor?

2.2. Objetivos de la investigación

- Objetivo General:
  - Explorar conjuntos de datos de la plataforma TripAdvisor con la intención de obtener información y descubrir patrones de comportamiento acerca de diversas variables que conforman el mercado hotelero.
- Objetivos Específicos:
  - Describir cómo algunas técnicas de minería de datos y estadísticas pueden ser aplicadas sobre los datos de Didi Soho Hotel y Blue Soho Hotel que provienen de la plataforma TripAdvisor;
  - Mostrar cómo algunas técnicas de minería de datos y estadísticas permiten obtener información general acerca de los hoteles 3 estrellas en Palermo Soho y ciertas particularidades de sus consumidores;
  - Aplicar programas informáticos que procesen los diferentes tipos de datos (numéricos, categóricos y textos) que existen en la plataforma TripAdvisor;
  - Enunciar algunas formas en que la información generada podría facilitar la toma de decisiones estratégicas y la gestión de los recursos de una empresa hotelera;
  - Utilizar la información y resultados expuestos para confeccionar gráficos que puedan ser interpretados con facilidad y/o contribuir con la toma de decisiones;

- Indicar posibles formas de mejorar las técnicas para poder ulteriormente obtener información más refinada y/o acertada, o bien, poder seguir sosteniendo o rechazar las hipótesis que fueron generadas;
- Proponer futuras líneas de investigación que podrían iniciarse dentro y/o fuera de las disciplinas involucradas y los ejes de la presente investigación.

2.3. Hipótesis

Al tratarse de un trabajo de investigación de naturaleza exploratoria, no necesariamente se formularán hipótesis, ya que no se podría presuponer (o afirmar) algo que apenas va a explorarse (Hernández et al., 2014, p. 116). Por otro lado, al tratarse además de una investigación del tipo mixta, gran cantidad de otras hipótesis emergerán y serán formuladas durante el trabajo para estudios futuros (Hernández et al., 2014, p. 576). Principalmente, debido a que la exploración y explotación de datos mediante diferentes programas de procesamiento de datos darán lugar a la generación de información (Hernández et al., 2014, p. 545), y por lo tanto, se generarán nuevos interrogantes que darán lugar a otras líneas de investigación en diversas disciplinas. Además, también se utilizarán hipótesis estadísticas a ser sometidas a distintas pruebas de hipótesis mediante herramientas de estadística inferencial (Hernández et al., 2014, p. 304).

2.4. Viabilidad

Primero que nada, para establecer la fuente de extracción de datos. ¿Por qué se consideró TripAdvisor como principal fuente de datos? Básicamente, porque *"TripAdvisor® es el sitio de viajes más grande del mundo… los sitios de la marca TripAdvisor conforman la comunidad de viajes más grande del mundo, con 390 millones de visitantes mensuales promedio exclusivos[1], y 435 millones de opiniones y comentarios sobre 6.8 millones de alojamientos, restaurantes y atracciones."* (TripAdvisor, Inc., 2017).

Con respecto al software principal de recolección, constitución, manipulación, análisis y visualización de datos. ¿Por qué considerar R como herramienta principal de procesamiento de datos? Simplemente porque "*R provee una amplia variedad de técnicas estadísticas y gráficas,*

[1] Fuente: registros de TripAdvisor, visitantes mensuales promedio exclusivos pertenecientes al tercer trimestre del año 2016.

y es altamente extensible. Una de las fortalezas de R es la facilidad con la que los gráficos bien diseñados y de calidad para la publicación pueden ser producidos, incluyendo símbolos matemáticos y fórmulas donde sea necesario. Gran atención se ha prestado a predeterminados de diseños menos seleccionados en gráficos, pero el usuario retiene el control total. R está disponible como un software gratuito bajo los términos de la Licencia General Pública del Free Software Foundation." (R Core Team, 2017).

Con respecto a la tipología de los datos que no son cuantitativos, es decir, los que son introducidos por el usuario mediante palabras, se tratan de datos de texto (cadena de caracteres) y están sujetos a los cánones de una lengua natural, que en este trabajo será la lengua española. Al estar contando con técnicas ampliamente aplicadas a la lengua inglesa, estas técnicas podrían adaptarse, aunque no del todo, a la lengua española.

2.5. Deficiencias

De acuerdo con un sin número de internautas: debido a las débiles barreras que TripAdvisor exige al usuario para publicar una opinión, es muy posible que cualquier internauta pueda redactar su opinión sin comprobar si su experiencia fue real o no. Por otro lado, de acuerdo con lo descrito en la página de esta plataforma, TripAdvisor filtra, modera, y detecta fraudes, entre otras, en las opiniones para garantizar que se cumplan las directrices de publicación (TripAdvisor, Inc., 2017).

Con respecto a otras fuentes de datos. ¿Por qué no se consideraron las opiniones de otras plataformas como Booking.com, Google Maps, Expedia.com, Trivago.com? Estas plataformas online presentan uno o más de los siguientes inconvenientes:

- Presionan mediante medios electrónicos al consumidor a opinar, dando también la posibilidad de omitir el texto, por lo que se puede dejar la opinión entera sin algunos componentes o bien el consumidor puede no estar de humor para comentar apropiadamente, llegando a emitir una opinión vaga o poco certera;
- Son inconvenientes a la hora de buscar las opiniones; ya que su propósito principal no es acumular opiniones de los consumidores, la llegada a estas por parte del usuario depende de un trayecto más largo, o bien, situado en una parte de la pantalla en donde el internauta llega a obviar más que otros sectores de la pantalla;
- El formato de la opinión dificulta la aplicación de técnicas de minería de datos para el sometimiento de las muestras a experimentos;

- Presentan dificultades o inconveniencias para el investigador de desarrollar una técnica de extracción de datos;
- No contienen métodos de filtro, moderación y/o detección de opiniones fraudulentas.

Con respecto al software utilizado para la mayoría de las tareas, R requiere dos tipos de habilidades principales, el primero, es la habilidad de programador, en donde uno debe ingeniárselas para escribir un código adaptado a los resultados que quiera lograr; el otro, es el conocimiento en materia estadística, no solamente en definiciones, fórmulas y funcionalidad, sino también su equivalente en inglés, ya que los manuales están escritos en dicho idioma.

Por otro lado, otras herramientas de procesamiento de datos, tan populares como R para las comunidades de profesionales en ciencia de datos, como Python, Matlab, SAS, SPSS, entre muchos otros, no fueron considerados para procesar los datos del siguiente trabajo, debido a uno o más de los siguientes motivos:

- El software no es gratuito;
- El software no ofrece una interfaz amigable;
- El software requiere conocimientos que el investigador no posee, o bien, el tiempo invertido en el aprendizaje es mayor;
- El software exige especificaciones al ordenador que actualmente no posee, y para tenerlas hay que invertir una significativa cantidad de dinero y/o tiempo;
- El software está perdiendo popularidad según las últimas tendencias, y por lo tanto, se prevé que quedará obsoleto, por lo que no es conveniente para proyectos futuros seguir utilizando la herramienta;
- El software no permite personalizaciones en la herramienta, por lo que habría que conformarse con funciones ya predefinidas;
- El software no permite la modificación y manipulación de los datos;
- El software no puede procesar cierto tipo de datos;
- El software no ofrece una amigable visualización de datos, por lo que no permitiría una fácil interpretación.

Con respecto a los datos recolectados, sobre todo aquellos no numéricos, tienden generalmente a presentar ciertos inconvenientes. Hohendahl (2011) aclara que:

> Si bien el estado del arte del PLN permite hoy procesar un texto en forma decente, éste estado del arte choca contra un patrón muy frecuente culturalmente cuando se trata de escribir texto: los errores de escritura de los que estamos hablando al amparo de la enorme variedad de texto viable, con inclusiones multiculturales, acrónimos, fórmulas y palabras prestadas de múltiples idiomas (‘loan words’) además de un sinnúmero de términos nuevos o parasintéticos de uso muy frecuente… Estos errores son muy difíciles de detectar y pueden ser de muchos tipos y de hecho los más frecuentes son los involuntarios y los que ocurren por falta de conocimiento o base cultural, como ser el saber si un determinado nombre propio se escribe con cierta combinación extraña de letras no frecuentes en nuestro idioma habitual.

Esto dificultó la tarea de agrupación y procesamiento de datos, la constitución de variables y sus categorías, la exposición de resultados y la confección de gráficos de representación.

MARCO TEÓRICO

3.1. Elementos de la hotelería

Definición de "hotel":

- La Ley de Regulación de Alojamientos Turísticos de la Ciudad Autónoma de Buenos Aires (Ley Nº 4.701, 2013), define que:

  > Un "Hotel" es un establecimiento que brinda servicio de alojamiento y otros complementarios, conforme a los requisitos que se indican para cada categoría, en habitaciones con baño privado y ocupa la totalidad o parte independiente de un inmueble, constituyendo sus servicios y dependencias un todo homogéneo.

- La Ley Nacional de Hotelería (Ley Nº 18.828, 1970), establece como sujeto a:

  > Los establecimientos comerciales en zonas turísticas o comprendidos en planes nacionales de promoción del turismo y los que por sus características el organismo de aplicación declare de interés para el turista, que ofrezcan normalmente hospedaje o alojamiento en habitaciones amuebladas por períodos no menores al de una pernoctación, a personas que no constituyan su domicilio permanente en ellos, quedan sujetos a la presente Ley y a las normas que se dicten en su consecuencia.

Definición de "servicio de alojamiento turístico" o "producto hotelero":

- La ley de Regulación de Alojamientos Turísticos de la Ciudad Autónoma de Buenos Aires (Ley Nº 4.701, 2013), Art. 3 menciona:

  > Servicio de alojamiento turístico. A los fines de esta ley se entiende por servicio de alojamiento turístico, aquél que se presta en establecimientos de uso público, en forma habitual o temporaria, por una tarifa y un período determinado, al que pueden sumarse otros servicios complementarios, siempre que las personas alojadas no constituyan domicilio permanente en ellos.

- Conde y Amaya (2007, p. 76) reconocen que:

  > El producto hotelero está formado por el conjunto de bienes y servicios que se ofrecen en el mercado, para el confort material y espiritual, en forma individual o en una gama muy amplia de combinaciones resultantes de las necesidades y deseos del consumidor al que le llamamos

> turista… El producto hotelero es la combinación de una serie de elementos tangibles e intangibles que solo afloran en el momento mismo de consumo.

- Kotler et al. (2011, p. 15) asocia la habitación del hotel como un producto tangible o algo que tiene propiedades físicas; además, forma parte de la oferta turística.

3.2. Elementos de la comunicación

Definición de "opinión":

- Según la RAE (Real Academia Española, 2014), una "opinión" (de raíz latina *opinio, -ōnis*) es definida como: "Juicio o valoración que se forma una persona respecto de algo o de alguien"; o bien, "fama o concepto en que se tiene a alguien o algo".

Acerca del "contenido generado por el usuario":

- El iBit (2011) entiende como UGC (User Generated Content, contenido generado por el usuario) a las *"informaciones publicadas en la red por sus usuarios"*.
- Basándose en una de las citas que hace Fernandez (2014; Curras et al., 2011), "el UGC (User Generated Content) es un tipo de información disponible a través de la red, sobre la cual los usuarios cada vez más exclusivos y exigentes basan buena parte de sus decisiones de compra".

Acerca de la "comunicación electrónica boca-oído" o "eWOM":

- Salvi et al. (2013; Litvin et al., 2008) definen eWOM como "todas las comunicaciones informales dirigidas a los consumidores mediante tecnologías basadas en Internet relacionadas con el uso o características de bienes y servicios, o de sus vendedores".
- Fernandez (2014; Hennig-Thurau et al., 2004, p.39) dice que el eWOM es "cualquier declaración positiva y negativa realizada por ex clientes, clientes actuales o clientes potenciales sobre un producto o empresa y que está a disposición de multitud de personas e instituciones vía internet".
- Berne (2015, p. 609; Negroponte & Maes, 1996; Dellarocas, 2003) menciona que "con el desarrollo de Internet, concretamente de la Web 2.0, los consumidores empezaron a emitir online sus opiniones. Esta nueva forma de comunicación, producida de manera íntegra por los usuarios, se denomina boca-oído electrónico o e-WOM".

3.3. Elementos de la teoría de la decisión

Acerca de lo que significa "decidir" o "tomar una "decisión":

- Según Bonatti et al. (2011, p. 21):

> Decidir es un proceso voluntario, sistemático, que a través de un análisis subjetivo, en ejercicio del razonamiento y con la emoción propia del ser humano, obtiene la elección/acción de una alternativa para cumplir con los fines, objetivos, propósitos previamente definidos, clarificados y ponderados por el sujeto que llamaremos decisor.

La "incertidumbre" y la "información" como conceptos antagónicos:

- Según Bonatti et al. (2011, p. 70):

> Para un decisor existe incertidumbre cuando en un momento determinado, encontrándose en posesión de cierto conocimiento, no sabe exactamente cuál es, cuál fue o cuál será el comportamiento del universo bajo análisis. Por otro lado, obtener información implica aprender y reducir la incertidumbre.

Acerca de los distintos niveles de incertidumbre:

- Según Bonatti et al. (2011, p. 76) existen 4 niveles de incertidumbre, estas son: Certeza, Riesgo, Incertidumbre y, el más importante para la investigación, Ambigüedad. En este último nivel, es casi imposible identificar las variables relevantes, no es posible contar con propensiones confiables y las estimaciones son poco aproximadas, la incertidumbre toma un rol protagónico. Aun así, se podrían realizar estudios utilizando diferentes herramientas como los modelos dinámicos no lineales, las analogías, y reconocimiento de pautas.

Acerca del concepto de "estrategia":

- La "estrategia" es un conjunto de cursos de acción afectados por distintos estados de variables no controlables generando resultados propios en un plazo determinado (Bonatti, 2011, p. 166).

3.4. Elementos del marketing

Acerca de la "información para el marketing":

- Kotler et al. (2011, p. 146) dice que:

> Para crear valor para los clientes y construir relaciones con ellos, las empresas necesitan obtener primero información actualizada sobre lo que sus consumidores necesitan y quieren. Las empresas utilizan esta visión de cliente para desarrollar ventajas competitivas. «En el hipercompetitivo mundo actual», sostiene un experto en marketing, «la carrera para obtener una ventaja competitiva es realmente una carrera para obtener una visión del mercado y de sus consumidores».

Acerca del "marketing" como decisión:

- Kotler (2011, p. 87) dice que *"la planificación de marketing implica la toma de decisiones sobre las estrategias de marketing que ayudarán a la empresa a alcanzar sus objetivos estratégicos generales… la lógica del marketing es crear valor para el cliente y lograr relaciones rentables."*

Acerca del "modelo de segmentación vincular":

- Wilensky (2006; Caden, 1986) explica acerca de este modelo confeccionado por Caden, el cual analiza y se focaliza únicamente en el vínculo que existe entre el consumidor y el producto/servicio. Básicamente, existen dos polos principales que son el "simbiótico" (caracterizado por lo afectivo) y el "discriminado" (caracterizado por lo racional), en donde el primero se trata de la unión entre el consumidor con el producto, y el segundo es la separación entre estos. Cada polo posee dos tipos de vínculos, que son los siguientes:
  - Vínculo comunitario: del polo simbiótico, responde al sentido de pertenencia, en donde se manifiestan valores como la lealtad, tradición, consenso, continuidad generacional;
  - Vínculo materno-filial: del polo simbiótico, responde al sentido de protección, en donde se manifiestan valores como la seguridad, afecto, nutrición, salud, y gratificación;
  - Vínculo simbologista: del polo discriminado, responde al sentido de identidad, en donde se manifiestan valores como el prestigio/status, estética/belleza, sensualidad refinada, saber de convenciones;

- Vínculo racionalista: del polo discriminado, responde al sentido de funcionalidad, en donde se manifiestan valores como practicidad, rendimiento, multifuncionamiento, saber, y precio.

3.5. Elementos de la minería de datos

Definición de la "Minería de Datos":

- Witter, Frank y Hall (2011, p. 4) explican que la cantidad de datos en el mundo se incrementa desmesuradamente, por lo que el uso de computadoras debe facilitar su manipulación. La explotación de datos debe aportar al hombre de negocios a identificar patrones que permitan tomar mejores decisiones, y así generar rentas para la empresa. [i]

Definición de la "Minería de Textos":

- Witter et al. (2011, p. 386) describen que así como la minería de datos busca identificar patrones en los datos, la minería de textos busca hacer lo mismo en los textos. Pero a diferencia de los datos que las empresas normalmente poseen en sus bases de datos, los textos son datos no estructurados, amorfos, y por lo tanto, difíciles de manejar. [i]

Sobre la "Minería de Opiniones" y el "Análisis del Sentimiento":

- Para Pang y Lee (2008, p. 8), el término "minería de opiniones" surge de la popularidad de las comunidades de usuarios asociadas con las búsquedas en la web y la adquisición de información. En cuanto al "análisis del sentimiento", se trata de procesar el lenguaje natural y extraer el tipo de sentimiento luego de evaluar el texto, y así, se clasifican las reviews de acuerdo con su polaridad (positiva o negativa). [i]
- Bing (2012, p. 7) dice que, ya el "análisis del sentimiento" es también llamado "minería de opiniones". Esta se trata de analizar las opiniones, sentimientos, evaluaciones, consideraciones, aptitudes y emociones de las personales con respecto a los productos, servicios, organizaciones, individuos, problemas, eventos, temas y sus atributos. [i]

Acerca de la "frecuencia de términos":

- Según Witter et al. (2011, p. 328) la "frecuencia de términos" se constituye con las "cadenas de caracteres", a las cuales se les puede atribuir un atributo numérico que señala qué tan usual se lo puede encontrar en el texto. Es normal que algunos términos

cambien de forma total o parcialmente para constituir el mismo grupo que otras cadenas de caracteres, y, por lo tanto, contar con el mismo atributo numérico. [i]

Acerca de la "extracción de aspectos":

- Según Bing (2012, p. 67), esta puede verse como una tarea de extracción de información. Sabiendo que una característica clave de una opinión es que esta siempre quiere comunicar sobre algo, y este algo suele ser un aspecto o tópico que puede ser extraído de la oración. Por ejemplo, en "el auto es caro", aunque "caro" es un término de sentimiento, esta oración también indica un "precio". [i]

Acerca de la "asociación de términos":

- Pecina (2009, p. 1) [i] explica que la asociación de términos o asociación léxica, se refiere a la asociación entre palabras del lenguaje natural. Existen tres tipos:
  - Asociación por colocación: establece la combinación de dos palabras para formar grupos de palabras (ejemplo: cirugía plástica, armas de destrucción masiva),
  - Asociación semántica: establece la relación semántica entre dos palabras (ejemplo: enfermedad – enfermo, perro – gato, conejo - zanahoria),
  - Asociación de lenguajes cruzados: establece como se traduciría una palabra (ejemplo: perro (en español) – dog (en inglés)).
- Para Silge y Robinson (2017), buena cantidad de análisis de textos son basados en la relación entre términos, examinando qué términos tienden a seguidamente suceder a otros, o bien, si entre los mismos documentos los términos tienden a coocurrir. [i]

Acerca de la "clasificación por subjetividad":

- La clasificación por subjetividad se trata de aquella técnica que clasifica un documento, de acuerdo con el vocabulario que utiliza, en "subjetivo" u "objetivo". Un texto objetivo describe y expresa información acerca de hechos, mientras que un texto subjetivo normalmente da puntos de vista y opiniones. De hecho, expresa evaluaciones, emociones, creencias, especulaciones, juicios de valor, estado de ánimo, etc. (Bing, 2012; Wiebe, Bruce y O'Hara, 1999). De acuerdo con una investigación de la conducta del consumidor, existen dos clasificaciones en cuanto a las evaluaciones que realizan los consumidores: "racional" y "emocional" (Bing, 2012, p. 28; Chaudhuri, 2006). [i]

Acerca de la "clasificación por análisis del sentimiento":

- Para Bing (2012, p. 62) la polaridad está compuesta por la naturaleza semántica de cada término, los cuales dejarían un saldo "positivo", "negativo", o "neutro". Esta técnica representa la orientación del sentimiento de los usuarios reflejado en el discurso utilizado. [i]

Acerca de las "reglas de asociación":

- Para Witter et al. (2011, p. 72)[i] las reglas de asociación son, además de reglas de clasificación, aquellas que predicen atributos y combinaciones de atributos, y generalmente expresa regularidades que subyacen en el conjunto de datos. Existen tres tipos de índices que contribuyen con la interpretación:
  - Confianza: esta indica la proporción entre la cantidad de casos favorables y casos posibles;
  - Soporte: indica la cantidad de casos en total que obedecen esta regla;
  - Lift: indica el incremento de casos para la regla.

3.6. Elementos de la estadística

Acerca de la "estadística"

- Capriglioni (2003a):

> Se llama ESTADÍSTICA a la disciplina científica que crea, desarrolla y aplica los adecuados métodos de recopilación de datos, y su evaluación, para transformarlos en informaciones con las cuales se describan objetivamente las distintas situaciones investigadas, se analice el comportamiento de determinadas características que poseen las UNIDADES EXPERIMENTALES, y se tomen decisiones en condición de incertidumbre.

Acerca de la "estadística descriptiva":

- Según describe Capriglioni (2003a, p. 41), el análisis descriptivo se utiliza al inicio de cualquier investigación, ya sea para establecer métodos de trabajo, controlar gestiones, describir relaciones y comportamientos de las variables, etc. Por otro lado, los entes que proporcionan los datos de la investigación, al ser medidos, observados o entrevistados, generalmente se encuentran en forma desordenada. Por lo que para obtener rápidamente

y en forma resumida, se han creado distintas medidas, cuyos fundamentos, utilización e interpretación, se explican mediante el análisis descriptivo de este autor.

Acerca de la "estadística inferencial":

- Capriglioni (2003b, p. 13) define que:

> Se llama "inferencia estadística" a cualquier afirmación que se realiza sobre una determinada población, basándose en los datos obtenidos con una muestra, pudiéndose obtener, a partir del cálculo de probabilidad, una determinada medida de la incertidumbre que se genera. Esto significa que se INFIERE la población a partir de la muestra. De esta manera, las conclusiones a que se llegan, por estar basadas en "ignorancias parciales", producen un cierto grado de duda.

Sobre la diferencia entre la estadística "Paramétrica" y "No Paramétrica":

- Según Anderson (2008, p.813), los métodos paramétricos son más potentes y refinados cuando las suposiciones necesarias sobre la distribución de probabilidad de la población son apropiadas. En los casos en que los datos son nominales y ordinales o en los casos en que las suposiciones requeridas por los métodos paramétricos son inapropiadas, sólo se cuenta con los métodos no paramétricos. Por lo que, se considera que tienen una aplicación más general que los métodos paramétricos.

Acerca de la estadística y la minería de datos:

- Han et al. (2012, p.23) explican que la estadística estudia la colección, análisis, interpretación o explicación, y presentación de los datos. Por lo que existe una conexión inherente entre la minería de datos y la estadística. La estadística es útil para minar varios patrones de conducta de los datos, así como también, entender los mecanismos subyacentes que generan y afectan los patrones. [i]

METODOLOGÍA

4.1. Tipología de la investigación

El diseño metodológico consta de un tipo de investigación exploratorio, en donde los datos recopilados fueron sometidos a distintas herramientas informáticas de procesamiento de datos, con el fin de examinar un tema poco investigado e indagarlo desde una nueva perspectiva (Hernández, 2014, p. 91). Por otro lado, tuvo un enfoque mixto en el sentido de que ambas metodologías cualitativas y cuantitativas fueron consideradas para este trabajo (Hernández, 2014, p. 533). El diseño es a su vez concurrente, ya que tanto los datos cuantitativos como los cualitativos se recolectan en un mismo momento. El diseño también sigue la triangulación, en el sentido que se alternan ambas metodologías, y además, queda a criterio del investigador la no corroboración o no confirmación del resultado de los datos (Hernández, 2014, p.557). La corroboración o confirmación fue expuesta para tratarse en futuras investigaciones.

4.2. Fuentes de datos y herramientas

Para la fuente de datos principal se utilizó la plataforma online TripAdvisor. Se recopilaron los datos mediante el lenguaje de programación R[2] y sus paquetes (conjunto de funciones Ad Hoc). Se obtuvieron los datos de los dos hoteles mediante una técnica llamada "web scraping", el cual toma los datos del código fuente de la totalidad de las páginas necesarias.

Para el procesamiento de datos se utilizó tanto el lenguaje R (y sus paquetes correspondientes según la herramienta a utilizar) como el software Excel (Microsoft – Versión Student 2013) junto a sus extensiones. Para su aplicación, fue necesaria una previa experimentación con la herramienta para poder ajustarla a las necesidades del investigador y los objetivos del presente trabajo. Esta etapa de preparación de las herramientas no será abordada en el presente trabajo, ya que no corresponde con el tipo de diseño ni con la formulación del problema. Sin embargo, esta etapa fue necesaria para contar con herramientas a medida lo suficientemente potentes para poder llevar a cabo la investigación propuesta.

Para la representación y visualización de datos se utilizó también R, Excel (Microsoft – Versión Student 2013), y en caso de necesitarse algún tipo de manipulación especial para mejorar la

[2] Paquetes principales utilizados: "rvest", "RSelenium", "XML", "RCurl".

interpretación (modificando las fuentes de texto, colores y nitidez de la imagen o gráfico), se utilizó Adobe Illustrator.

4.3. Diseño muestral

4.3.1 Población.

La población se estableció como todo huésped que alguna vez se haya hospedado en algún hotel de gama media que se sitúan en la zona de Palermo Soho[3].

Para realizar el estudio, se optó por reducir el muestreo a una sola plataforma online: TripAdvisor, y a dos hoteles en particular. Siendo uno el hotel principal (sometido a todas las técnicas) y el otro el hotel secundario o de comparación (sometido a unas pocas técnicas). Para esto, se establecieron los siguientes filtros en el buscador de la plataforma[4]:

CAMPO	OPCIÓN	CONSIDERACIONES
Encontrá:	Hoteles	Al ser una empresa turística, se ajusta a los objetivos de la presente investigación.
Cerca:	Distrito Capital Federal, Argentina	Siendo la Ciudad Autónoma de Buenos Aires un importante destino del turismo receptivo se optó por esta ubicación.
Categoría del hotel	3 estrellas (★★★☆☆)	Se optó por un hotel de gama media con la intención de evitar los extremos, ya que se previó que dos hoteles competidores de 4 o 5 estrellas no posean diferencias significativas, y dos hoteles de 1 o 2 estrellas tengan tantas falencias que quizás los datos solo den lugar a información muy evidente.
Vecindarios	Palermo	Por ser un barrio en donde se encuentra la mayor concentración de hoteles de mediana categoría en la Ciudad de Buenos Aires (Dirección General de Estadísticas y Censos, 2016).
Cadena hotelera	Hoteles independientes	La intención es descartar las cadenas hoteleras que estas ya poseen un branding específico y una reputación que condiciona las opiniones y dificulta la comparación con otros hoteles.
Precio por noche[5]	De $576 a $2476 pesos la noche[6]	La intención es establecer un rango aceptable para el turista.

[3] Nombre otorgado por denominación popular a un sub-barrio no oficial que se encuentra en el barrio de Palermo de la Ciudad Autónoma de Buenos Aires.

[4] Especificaciones de la interfaz del usuario. País: Argentina; idioma: español; usuario: ninguno.

[5] Conversión al cambio del día 01 de marzo del año 2017, según DolarHoy.com. Valor del dólar establecido mediante el promedio entre las cotizaciones promedio (compra: $15.35 y venta: $15.80), siendo la media de $15.575 pesos argentinos.

[6] Rango establecido de $37 a $159 dólares la noche por la página PriceofTravel el día 01 de marzo de 2017. https://www.priceoftravel.com/10/argentina/buenos-aires-prices.

Datos tomados de TripAdvisor el día 01 de marzo del 2017[7]

Una vez establecidos los filtros, se percibe una significativa similitud entre los siguientes dos hoteles, mediante las siguientes comparaciones:

CAMPO	HOTELES EN CUESTIÓN		OBSERVACIONES
Nombre	Didi Soho Hotel	Blue Soho Hotel	Poseen nombres similares
Ubicación	Honduras 4762	El Salvador 4735	Se encuentran a solamente una cuadra de distancia
Clasificación Orgánica	Ambos aparecen consecutivamente en los puestos 25 y 26, primera página.		
Clasificación por listado de hoteles	387 de 450 hoteles en Buenos Aires	386 de 450 hoteles en Buenos Aires	Poseen una clasificación general similar
Tarifas (precio de la habitación por noche)[8]	$950 por Despegar.com $968 por Booking.com $970 por Expedia.com	$944 por Booking.com	Ambos hoteles poseen tarifas similares en cuanto al precio
Puntaje general	3 estrellas y 1/2		Ambos hoteles poseen la misma puntuación
Cantidad de Opiniones	319 opiniones	277 opiniones	Cantidades de opiniones no difieren en más del 15%
Cantidad en español Cantidad en portugués Cantidad en inglés	192 83 81	151 90 73	
“Explorar hoteles similares”	Blue Soho Hotel en primer puesto	Didi Soho Hotel en primer puesto	La página los muestra como “hoteles similares” entre sí
“Hoteles que posiblemente te gusten…”			

Datos tomados de TripAdvisor el día 01 de marzo del 2017[9]

De esta forma, se estableció tanto Didi Soho Hotel (como hotel principal, objeto de estudio) y Blue Soho Hotel (como hotel secundario, competencia de Didi Soho Hotel) para realizar la exploración y explotación de datos.

4.3.2. Muestreo.

En la etapa de muestreo, se tomaron los siguientes campos para constituir los conjuntos de datos que fueron sometidos a los diferentes experimentos:

- “Nivel de crítico”;

[7] Ver capturas de pantalla de los hoteles en la sección Anexos.

[8] Tarifas tomadas el día 01 de marzo del 2017, para los días 04 de marzo al 07 de marzo del mismo año, siendo el rango de fechas establecido por defecto por TripAdvisor.

[9] Ver capturas de pantalla de los hoteles en la sección Anexos.

- “Título de la opinión”;
- “Texto de la opinión”;
- “Puntaje otorgado”;
- “Fecha del alojamiento”;
- “Propósito del viaje”;
- “Dispositivo móvil”.

Por otro lado, existen ciertos que campos no fueron recolectados ni examinados debido a que no se previó que fueran potencialmente útiles para la presente investigación. A continuación, se describen los campos que no constituyeron el conjunto de datos y sus motivos:

- “Nombre del usuario”: al consistir en el primer nombre de la persona y la inicial de su primer apellido, se prefirió evitar ambigüedades en los nombres y la repetición de las mismas en la población. En ciertos casos, se les asignó temporalmente un número que indica el orden de aparición en la plataforma;
- “Procedencia del usuario”: debido a que no todos los usuarios especifican este campo, se decidió, por motivos de inconsistencia, descartar este atributo de la constitución del conjunto de datos;
- “Cantidad de opiniones”, “Cantidad de opiniones sobre hoteles”, y “Cantidad de votos útiles”: estos campos resultarían poco útiles, ya que estos afectan directamente a la variable “nivel de crítico”, por lo tanto, esta última se utilizará para la investigación;
- “Fecha de escritura de la opinión”: no será considerada ya que se consideró más relevante la fecha en la cual efectivamente el huésped se alojó en el establecimiento.
- Puntajes individuales en “ubicación”, “limpieza”, “servicio”, “relación calidad-precio”, “habitaciones”, “calidad de descanso”: no se tomaron en cuenta, ya que esta no todos los usuarios la califican. Además, solo se pueden calificar tres de ellos y los cuales aparecen de manera aleatoria, sin posibilidad por parte del usuario de elegirlos;
- “Fotos subidas”: la cantidad y calidad de las mismas tampoco fueron tenidas en cuenta ya que no todos los usuarios comparten fotos, y además, las herramientas no fueron lo suficientemente poderosas para analizar archivos del tipo imagen;
- “Consejo sobre las habitaciones”: si bien podría llegar a brindar información útil, no todos los usuarios incluyen esta sección en sus opiniones; por lo que, se decidió por cuestiones de comodidad y consistencia, que esta variable debió ser descartada para constituir la base de datos;

- "Respuesta del representante del hotel": el texto mediante el cual el usuario representante del hotel utiliza para responderle al usuario resultó irrelevante para este trabajo ya que, el administrador no siempre responde a todas las opiniones, y además, esto no da información acerca del turista.

4.3.3. Criterios de muestreo.

El muestreo fue principalmente determinado por un tipo de muestreo intencional (Capriglioni, 2003b, p. 14), mediante los siguientes criterios:

- Las opiniones debieron estar escritas en español;
- El tamaño de la muestra debe ser múltiplo de 10, ya que TripAdvisor muestra cada 10 opiniones;
- Las opiniones que primero aparezcan en la descripción del hotel son más relevantes ya que serán las que primero verán los consumidores en proceso de decisión de compra; por este motivo, se dará prioridad al muestreo a este tipo de opiniones;
- La muestra debería ser lo suficientemente grande para representar la realidad en la empresa hotelera y para que el procesamiento de datos pueda encontrar patrones similares entre las opiniones;
- La muestra debería considerar un rango de tiempo significativo, abarcando todas las temporadas por una mínima cantidad de años relevante.

Se estableció que cada muestra conformada por las opiniones de un hotel específico tuviera una cantidad exacta de 100 opiniones (conjunto de unidades experimentales), las cuales fueron tomadas de acuerdo con el orden de aparición en TripAdvisor, aclarando nuevamente que se desconoce el algoritmo que ordena la aparición de las opiniones.

4.3.4. Constitución del conjunto de datos.

Los datos recolectados fueron almacenados en un archivo .csv con el propósito de poder ser invocados y utilizados por las distintas técnicas de procesamiento de datos de forma efectiva e inmediata. Aquí, los campos fueron constituidos por las poblaciones mencionadas anteriormente. A continuación, se detallan los posibles valores asignados para estas variables:

- "Número de usuario": van del número 1, y se incrementaron en 1, hasta coincidir con el tamaño de muestra (n). Esta variable fue constituida en escala discreta decreciente, siendo el "1" el valor más alto, y el valor de n el valor más bajo, ya que representa el

orden en el cual otros usuarios hubieran leído las opiniones, y por lo tanto, mayores probabilidades tuvieron de ser tenidos en cuenta cuando este nuevo usuario quiera tomar una decisión de compra;

- "Nivel de crítico": van del 0 al 6, siendo el "0" el valor más bajo, y el "6" el valor más alto, constituyendo así una variable en escala ordinal;
- "Título de la opinión" y "Texto de la opinión": varía según la opinión del usuario y fueron tratadas como variables del tipo cadena de caracteres para ser sometidas a experimentos de minería de datos.
- "Puntaje otorgado": van del "1" al "5", siendo el "1" el puntaje más bajo, y el "5" el puntaje más alto, constituyendo así una variable en escala discreta;
- "Fecha del alojamiento": de esta se distinguieron tanto el mes como el año para cada unidad. En muy pocos casos, esta fecha pudo no estar expresada por el usuario, y en este caso el valor asignado fue de "0". La combinación del mes y el año hizo que esta variable sea del tipo ordinal. Por otro lado, los valores de esta variable también fueron subdivididos en dos partes:
  - mes: se le asignó un valor numérico[10];
  - año: variable numérica en escala ordinal.
- "Propósito del viaje": de esta variable en escala nominal se distinguen 6 categorías diferentes, que fueron las siguientes:
  - Viaje en familia;
  - Viaje en pareja;
  - Viaje solitario;
  - Viaje de negocios;
  - Viaje con amigos;
  - No especificado: esta categoría se utilizó para el conteo de la muestra, es decir, controlar si la muestra llega a las 100 unidades; por lo que de ninguna forma se consideró, a esta categoría, como un segmento del mercado turístico.
- "Dispositivo móvil": es una variable booleana o variable dicotómica categórica, simplemente señala si el comentario fue escrito desde una Smartphone o no. Por lo que,

[10] El número "1" equivale al mes de "enero", el número "2" equivale al mes de "febrero", y así sucesivamente hasta llegar al número "12".

si el valor es “1”, el comentario fue escrito a través de un dispositivo móvil, y si el valor es “0”, el comentario no fue escrito a través de un dispositivo móvil.

4.4. Metodología aplicada

La metodología aplicada, junto con las pautas y consideraciones de trabajo, se constituyeron basándose en ciertos criterios, los cuales son tratados en este punto.

Primero, es propio del presente trabajo realizar una investigación acorde a principios científicos. Se estuvo estudiando un fenómeno específico y se utilizó tanto el marco teórico y métodos de otras disciplinas, por lo que se evidencia un papel interdisciplinar en el mundo académico del turismo como disciplina (Jafari, 2005, p. 50).

Segundo, es importante aclarar que al tratarse de un tipo de investigación exploratorio en donde se desconoce en términos prácticos estudios previos y evidencias en la utilidad del eWOM, esto dio lugar a una categoría de incertidumbre llamada “ambigüedad”, en donde falta información, o la misma es confusa, el contexto es altamente turbulento, y la identificación de las variables relevantes casi imposible. Ulteriormente, se espera que esta situación de ambigüedad debería transformarse en situaciones de riesgo o incertidumbre (Bonatti et al, 2011, p. 79). Esto se lo determinó, de acuerdo con los siguientes supuestos propuestos por el investigador:

- Se desconoce si los datos son los suficientemente consistentes como para brindar algún tipo de información;
- Se desconoce si las herramientas son lo suficientemente potentes para procesar todos los tipos de datos recolectados;
- Se desconoce si generar modelos de decisión a partir de la información a generar puede llegar a impactar positivamente al negocio;
- Se desconoce si el tipo de información que se puede generar resultaría ser un tipo de información innecesaria, confusa o que pueda derivar en la toma de una decisión que perjudique al negocio.

Tercero, conforme a la formulación del tema y los objetivos establecidos, los cuales se basan mayoritariamente en las ciencias económicas, se determinó que el investigador debe orientar el trabajo con la intención de responder a las necesidades del órgano gestor y decisor de una empresa turística. De esta forma, el análisis exploratorio y la explotación de los datos se

realizarán con la intención de aprender y obtener información para reducir la incertidumbre (Bonatti et al., 2011, p. 70), y así beneficiar el negocio hotelero. Además, se revisó anteriormente cómo ciertos autores como Kotler et al. (2011, p. 137) remarca la necesidad de obtener información acerca de los consumidores para tomar decisiones; y Levy (2012, p. 61) señala el estudio del CONES (conjunto esperado) y la identificación de segmentos para conformar un conjunto de información estratégica.

A continuación, el siguiente cuatro representa las técnicas y softwares que se utilizaron ordenados por técnicas, los cuales representan su respectiva parte:

<table>
<tr><th>Capítulos</th><th>Requerimientos[11]</th><th>Formas de representación</th><th>Tipo de razonamiento</th><th>Muestras</th><th>Herramienta utilizada</th></tr>
<tr><td>1) Estadística descriptiva</td><td></td><td>Gráfico de barras, cuadro comparativo.</td><td rowspan="2">Inductivo</td><td rowspan="2">Didi Soho Hotel y Blue Soho Hotel</td><td rowspan="4">R</td></tr>
<tr><td>2) Frecuencia de términos</td><td>1)</td><td>Nube de términos.</td></tr>
<tr><td>3) Asociación de términos</td><td>1), 2)</td><td>Gráfico de nodos, dendrogramas.</td><td rowspan="2">Deductivo</td><td rowspan="6">Didi Soho Hotel</td></tr>
<tr><td rowspan="2">4) Análisis del sentimiento</td><td rowspan="4">1), 2), 3)</td><td>Matriz de correlación</td></tr>
<tr><td rowspan="3">Inductivo</td><td rowspan="2">MeaningCloud</td></tr>
<tr><td>5) Clasificación por subjetividad</td><td>Gráfico de barras</td></tr>
<tr><td>6) Generación de tópicos</td><td>Tablas de contingencia</td><td rowspan="3">R</td></tr>
<tr><td rowspan="2">7) Reglas de asociación</td><td rowspan="2">1), 4), 5), y 6)</td><td>Gráfico de barras, tabla de valores.</td></tr>
<tr><td>Diagrama de dispersión</td><td rowspan="2">Deductivo</td></tr>
<tr><td>8) Estadística inferencial</td><td>1), 4), métodos de:
a) Wilcoxon;
b) Aleatoriedad;
c) Spearman;
d) Mann-Whitney.</td><td>Diagrama de dispersión, diagrama de cajas.</td><td>Inductivo y Deductivo</td><td>Didi Soho Hotel y Blue Soho Hotel</td></tr>
<tr><td>Matriz de correlación.</td></tr>
</table>

Fuente: cuadro comparativo de elaboración propia.

[11] Esta columna menciona los capítulos del trabajo anteriores y/u otras técnicas para llevar a cabo el desarrollo del capítulo correspondiente.

DESARROLLO DEL TRABAJO

5.1. Capítulo 1: Estadística descriptiva

Para esta primera técnica, no se utilizaron los datos recopilados por nuestra herramienta de muestreo, sino que se utilizaron los datos que la plataforma TripAdvisor otorga a primera vista sobre las opiniones de cada hotel. Que, al tratarse de una cantidad de datos relativamente pequeña, y al no presentar dificultades al tomarlas, su extracción se hizo de forma manual.

Para este primer abordaje, los datos tomados fueron de los hoteles Didi Soho Hotel y Blue Soho Hotel, con la intención de compararlos de a pares, obtener información de estos, tanto para favorecer el proceso de toma de decisiones como para generar varias hipótesis acerca de las poblaciones a partir de la utilización de las lógicas inductiva[12] y deductiva. Estos datos tomados se trataron específicamente de las frecuencias absolutas de la "calificación", "tipo de viajero" y "época del año", cada una de ellas filtradas por el "idioma" (español, inglés y portugués) que están explícitas en la plataforma.

En cuanto a técnicas de estadística descriptiva, fueron utilizadas las siguientes:

- Frecuencia absoluta y modo/moda (para las "calificaciones", "tipo de viajero", y "época del año");
- Media aritmética, mediana, coeficiente de asimetría, coeficiente de curtosis (para las "calificaciones") calculados mediante el lenguaje R[13].

Para la representación de los datos, se utilizaron las herramientas gráficas que ofrece Excel (2016 – Student Version), principalmente gráfico de barras y gráfico de torta o circular de acuerdo con el tipo de filtro, y se los analizó de a pares, comparando los datos de los dos hoteles en cuestión. A continuación, se presentan entonces las comparaciones mediante los gráficos

[12] Estas hipótesis serán creadas mediante una lógica inductiva, la cual se sabe que este tipo de razonamiento es refutado y/o su validación se ve altamente en riesgo de acuerdo con ciertos epistemólogos. Siendo Karl Popper (1902 - 1994) el más importante con su corriente llamada "Falsacionismo", seguido de su discípulo Imre Lakatos con el "Falsacionismo sofisticado".

[13] Principales paquetes utilizados: "Hmisc" y "fBasics".

obtenidos a partir de estos datos de la plataforma, seguidos de sus respectos análisis de información y conclusiones.

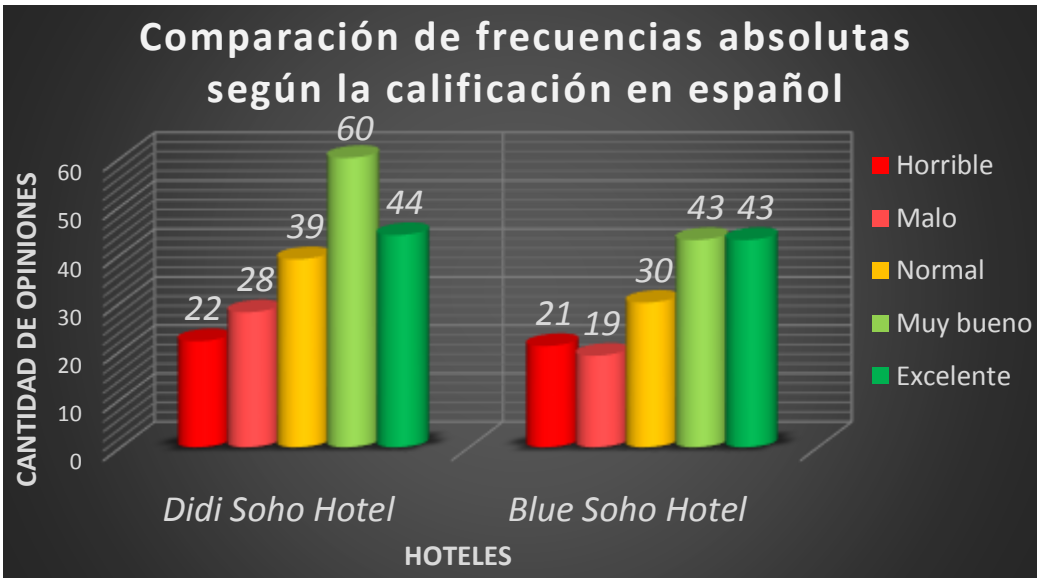
5.1.1. Comparación de hoteles según la calificación del usuario.

Debido a que la "calificación" es una variable aleatoria discreta, se procedió con calcular las medidas de tendencia central (media aritmética, mediana y moda), medidas de variación (desvío estándar), y medidas de concentración (coeficientes de asimetría y curtosis). En cuanto a las calificaciones, TripAdvisor muestra que la escala de calificaciones (1 a 5 puntos) es equivalente a la escala utilizada (horrible-excelente). Por lo que que consideró que el puntaje 1 equivale a "horrible", el puntaje 2 a "malo", y así, sucesivamente hasta llegar al 5 que equivale a "excelente". De esta forma se obtuvieron los siguientes gráficos, junto con sus respectivos análisis de la información obtenida separadas por el idioma utilizado, en conjunto con la creación de hipótesis a ser probadas en otras investigaciones:

5.1.1.1. En idioma español.

Mediante la utilización de los datos correspondientes a lo previamente explicado, se ha logrado confeccionar el siguiente gráfico:

Figura 1.1. Comparación de frecuencias absolutas según la calificación en español



Fuente: elaboración propia en base a datos de TripAdvisor.

Figura 1.2. Comparación de medidas descriptivas para la calificación en español

	Didi Soho Hotel	Blue Soho Hotel	Detalles
Media poblacional	3.393782	3.435897	

Desvío estándar	1.295168	1.363958	
Mediana	4	4	
Moda	"muy bueno" (60)	"muy bueno" y "excelente" (43)	Didi Soho Hotel solo tiene un pico en "muy bueno", mientras que Blue Soho Hotel tiene una distribución bimodal con dos picos ("horrible" y "muy bueno") y un valle en "malo"
Coeficiente de asimetría	-0.4394459	-0.4799968	Ambos poseen distribución asimétrica hacia la izquierda
Coeficiente de curtosis	-0.9193647	-0.9942921	Ambos poseen distribución platicúrtica

Fuente: elaboración propia en base a datos de R.

En estas dos poblaciones de opiniones escritas en idioma español se pudo observar que:

- Ambos poseen la misma mediana;
- La moda en Blue Soho Hotel está compartida entre los dos valores de orden mayor, a diferencia de Didi Soho Hotel en donde la moda está en "muy bueno";
- La media poblacional de Blue Soho Hotel es ligeramente superior (alrededor del 1%) al de Didi Soho Hotel;
- El desvío estándar de Blue Soho Hotel es ligeramente superior (alrededor del 5%) al de Didi Soho Hotel;
- Blue Soho Hotel posee un coeficiente de asimetría menor al de Didi Soho Hotel, por lo que los valores de Blue Soho Hotel a la izquierda están más dispersas;
- Blue Soho Hotel posee un coeficiente de curtosis menor al de Didi Soho Hotel, por lo que los valores de Blue Soho Hotel están menos concentradas en el pico de la distribución.

Por lo tanto, fue posible concluir que Blue Soho Hotel es ligeramente mejor calificado por los viajeros de habla española que Didi Soho Hotel, sin embargo, las medidas de variación y de concentración de la población muestran que las calificaciones son ligeramente menos tendientes a caer en el valor de la media aritmética en Blue Soho Hotel que en Didi Soho Hotel.

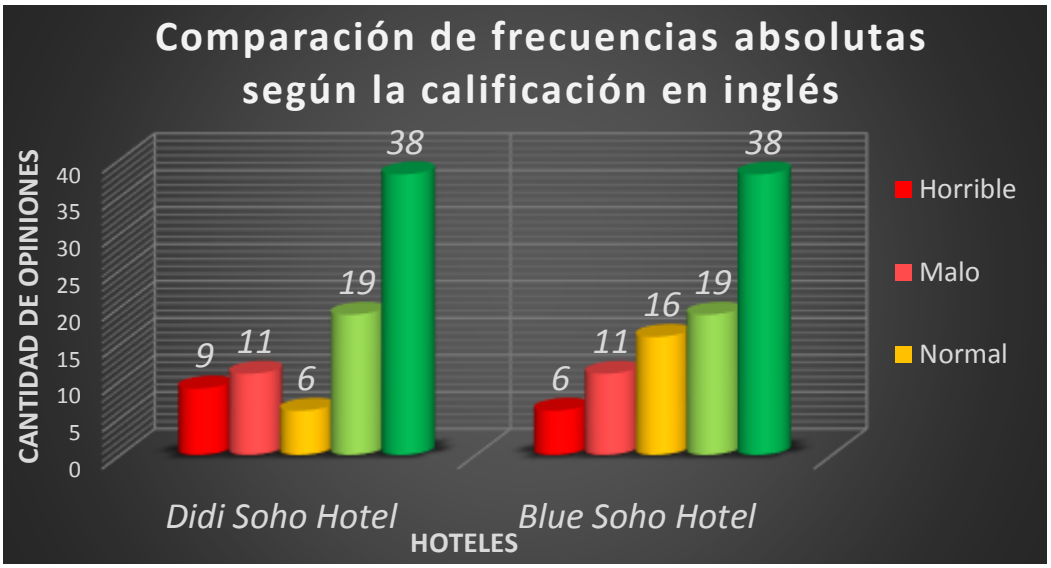
De esta forma se podrían constituir las siguientes hipótesis:

1. *"Los hispanohablantes perciben que el producto hotelero de Blue Soho Hotel es mejor al de Didi Soho Hotel";*
2. *"Los hispanohablantes poseen cierta tendencia a calificar en 4 puntos sobre 5 cuando se trata de hoteles 3 estrellas situados en el barrio de Palermo".*

5.1.1.2. En idioma inglés.

Mediante la utilización de los datos correspondientes a lo previamente explicado, se ha logrado confeccionar el siguiente gráfico:

Figura 1.3. Comparación de frecuencias absolutas según la calificación en inglés



Fuente: elaboración propia en base a datos de TripAdvisor.

También, fue posible confeccionar el siguiente cuadro comparativo en materia de estadística descriptiva:

Figura 1.4. Comparación de medidas descriptivas para la calificación en inglés

	Didi Soho Hotel	Blue Soho Hotel	Detalles
Media poblacional	3.795181	3.8	
Desvío estándar	1.420744	1.291285	
Mediana	4	4	
Moda	"excelente" (38)	"excelente" (38)	Didi Soho Hotel tiene un "valle" en "normal"
Coeficiente de asimetría	-0.8258866	-0.7133873	Ambos son asimétricos hacia la izquierda
Coeficiente de curtosis	-0.7936375	-0.734548	Ambos son platicúrticos

Fuente: elaboración propia en base a datos de R.

En estas dos poblaciones de opiniones escritas en idioma inglés se pudo observar que:

- Ambas poblaciones poseen casi la misma media poblacional (solo un 0,12% de diferencia);
- Ambas poblaciones poseen la misma mediana y moda;

- Didi Soho Hotel posee una desviación estándar mayor al de Blue Soho Hotel (un poco más del 9% mayor);
- Didi Soho Hotel posee una población mayormente dispersa en la izquierda de la distribución según el coeficiente de asimetría;
- Didi Soho Hotel posee una población mayormente platicúrtica que la de Blue Soho Hotel, y por lo tanto menos empinada que esta.

Así, fue posible concluir que, para las poblaciones de habla inglesa, Blue Soho Hotel es también ligeramente mejor calificado por los viajeros que Didi Soho Hotel.

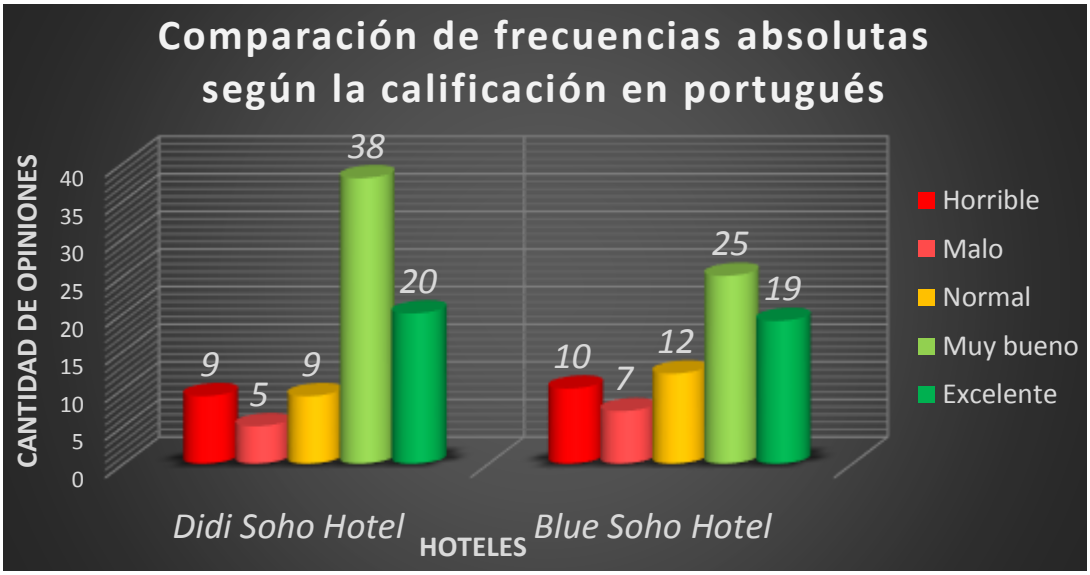
De esta forma se podrían constituir las siguientes hipótesis:

3. *"Los angloparlantes perciben que el producto hotelero de Blue Soho Hotel es mejor al de Didi Soho Hotel";*
4. *"Los angloparlantes poseen cierta tendencia a calificar en 5 puntos sobre 5 cuando se trata de hoteles 3 estrellas situados en el barrio de Palermo";*
5. *"Los angloparlantes poseen una percepción distinta a la de los hispanohablantes".*

5.1.1.3. En idioma portugués.

Mediante la utilización de los datos correspondientes a lo previamente explicado, se ha logrado confeccionar el siguiente gráfico:

Figura 1.5. Comparación de frecuencias absolutas según la calificación en portugués



Fuente: elaboración propia en base a datos de TripAdvisor.

También, fue posible confeccionar el siguiente cuadro comparativo en materia de estadística descriptiva:

Figura 1.6. Comparación de medidas descriptivas para la calificación en portugués

	Didi Soho Hotel	Blue Soho Hotel	Detalles
Media poblacional	3.679012	3.493151	
Desvío estándar	1.233158	1.344983	
Mediana	4	4	
Moda	"muy bueno" (38)	"muy bueno" (25)	Ambos tienen un valle en "malo"
Coeficiente de asimetría	-1.001832	-0.6274821	Ambas distribuciones son asimétricas hacia la izquierda
Coeficiente de curtosis	0.0467932	-0.8170646	La distribución de Didi Soho Hotel es leptocúrtica, pero casi mesocúrtico; y la distribución de Blue Soho Hotel es platicúrtica.

Fuente: elaboración propia en base a datos de R.

En estas dos poblaciones de opiniones escritas en idioma portugués se pudo observar que:

- La media poblacional de Didi Soho Hotel es ligeramente mayor a la de Blue Soho Hotel (aproximadamente un 5%);
- Tienen la misma mediana y moda (aunque Didi Soho Hotel posee más repeticiones);
- Blue Soho Hotel posee un desvío mayor al de Didi Soho Hotel (un poco más del 8% mayor);
- El coeficiente de asimetría dice que en Didi Soho Hotel hay más valores más dispersos por debajo de la media;
- El coeficiente de curtosis dice que en Didi Soho Hotel, las medidas de tendencia central están más concentradas en ese punto de lo que están para Blue Soho Hotel.

Así, fue posible concluir que, para las poblaciones de habla portuguesa, Didi Soho Hotel es también ligeramente mejor calificado por los viajeros que Blue Soho Hotel.

De esta forma se podrían constituir las siguientes hipótesis:

6. *"Los lusoparlantes perciben que el producto hotelero de Didi Soho Hotel es mejor al de Blue Soho Hotel";*
7. *"Los lusoparlantes poseen una mayor tendencia a calificar en 4 puntos sobre 5 cuando se trata de hoteles 3 estrellas situados en el barrio de Palermo a comparación del hispanohablante";*

8. *"Los lusoparlantes poseen una percepción distinta a la del hispanohablante y a la del angloparlante".*

5.1.2. Comparación de hoteles según el tipo de viajero.

5.1.2.1. En idioma español.

Mediante la utilización de los datos correspondientes a lo previamente explicado, se ha logrado confeccionar el siguiente gráfico:

Figura 1.7. Comparación de frecuencias absolutas según el motivo de viaje en español

Fuente: elaboración propia en base a datos de TripAdvisor.

En estas dos poblaciones de opiniones escritas en idioma español se pudo observar que:

- Para ambos hoteles el motivo de viaje que prevalece es el de "pareja";
- Para Blue Soho Hotel, existe una prevalencia aún mayor de huéspedes que viajaron en "pareja" a comparación del resto de los motivos de viaje;
- Según el total, los viajeros "solitarios" fueron los que menos hicieron presencia, pero aun así, en Blue Soho Hotel hubieron menos viajeros cuyo motivo fue el de "negocios";
- Para ambos hoteles, luego de los viajeros en "pareja", les siguen los viajeros en "familia", y luego por "amigos".

Entonces, fue posible constituir las siguientes hipótesis:

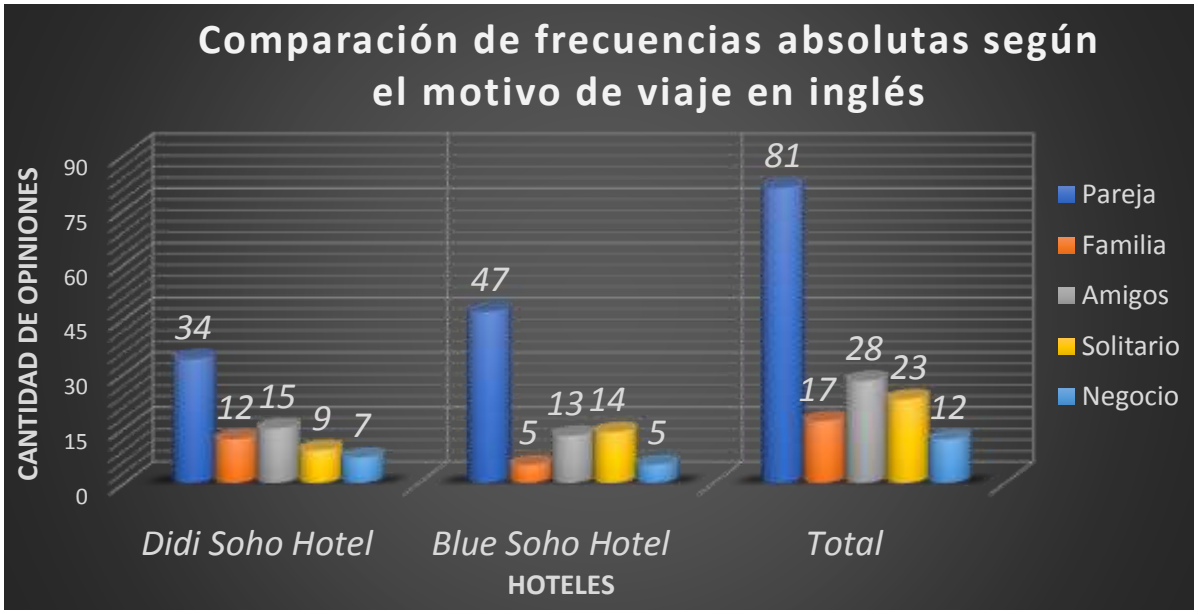
9. *"La gran mayoría de los turistas hispanohablantes que se hospedan en hoteles 3 estrellas de Palermo vienen en pareja";*

10. *"El segundo motivo por el cual los viajeros de habla hispana deciden hospedarse en hoteles 3 estrellas de Palermo es por viajes en familia;*
11. *"El tercer motivo por el cual los viajeros de habla hispana deciden hospedarse en hoteles 3 estrellas de Palermo es por viajes con amigos;*
12. *"El target principal de consumidores de Blue Soho Hotel es el viajero en pareja";*
13. *"Didi Soho Hotel no posee un claro target principal de viajero".*

5.1.2.2. En idioma inglés.

Mediante la utilización de los datos correspondientes a lo previamente explicado, se ha logrado confeccionar el siguiente gráfico:

Figura 1.8. Comparación de frecuencias absolutas según el motivo de viaje en inglés



Fuente: elaboración propia en base a datos de TripAdvisor.

En estas dos poblaciones de opiniones escritas en idioma inglés se pudo observar que:

- Para ambos hoteles el motivo de viaje que prevalece es el de "pareja";
- Para Blue Soho Hotel, existe una prevalencia ligeramente mayor de huéspedes que viajaron en "pareja" a comparación del resto de los motivos de viaje;
- A diferencia de los viajeros hispanohablantes, los viajeros en "familia" hicieron menos presencia que los que viajaron con "amigos";
- Para Didi Soho Hotel, a diferencia de las opiniones en español, y sabiendo que prevalece el motivo de viaje en pareja, el resto de los motivos de los poseen cantidades más próximas entre ellas.

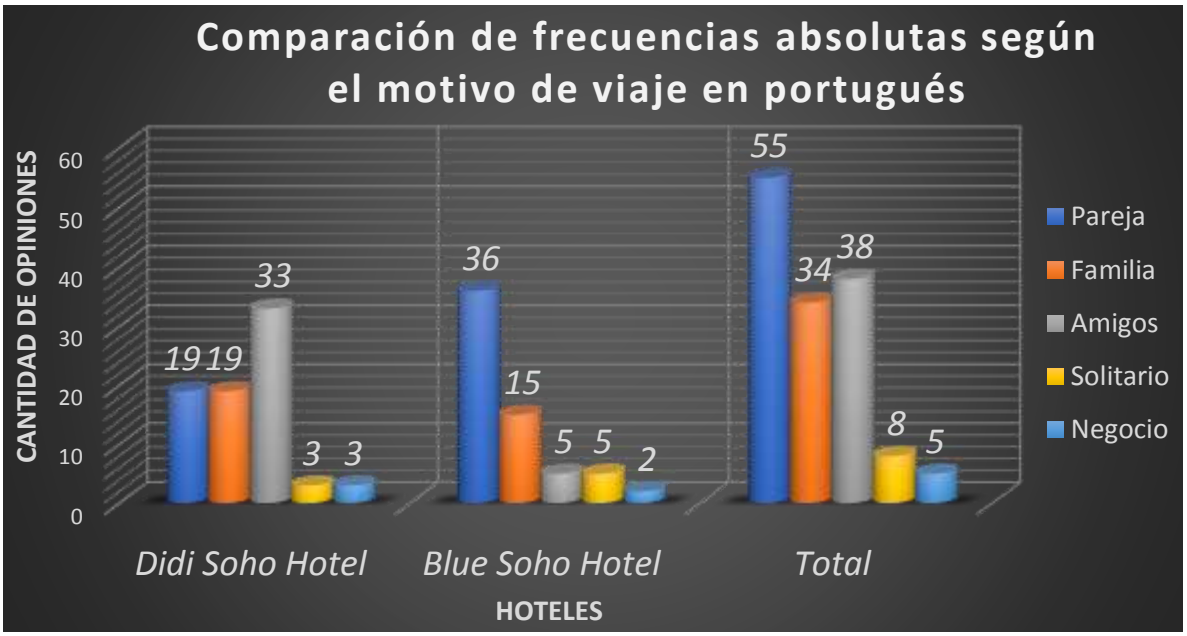
Así, fue posible constituir las siguientes hipótesis:

14. *"La gran mayoría de los turistas angloparlantes que se hospedan en hoteles 3 estrellas de Palermo vienen en pareja";*
15. *"Ambos hoteles compiten por el mismo target de angloparlantes que se hospedan en hoteles 3 estrellas de Palermo y cuyo motivo de viaje es ir en pareja";*
16. *"Ambos hoteles compiten menos por los viajeros angloparlantes cuyos motivos sean familia, amigos, negocio y solitario".*

5.1.2.3. En idioma portugués.

Mediante la utilización de los datos correspondientes a lo previamente explicado, se ha logrado confeccionar el siguiente gráfico:

Figura 1.9. Comparación de frecuencias absolutas según el motivo de viaje en portugués



Fuente: elaboración propia en base a datos de TripAdvisor.

En estas dos poblaciones de opiniones escritas en idioma portugués se pudo observar que:

- Para Didi Soho Hotel, existe una mayor cantidad de opiniones escritas por viajeros cuyo motivo fue ir con "amigos", mientras que para Blue Soho Hotel, sigue prevaleciendo el motivo de viaje en "pareja";
- Para Didi Soho Hotel, el segundo motivo de viaje más común entre los lusoparlantes es en "familia" y con "amigos" por igual; y el tercer motivo de viaje más común para los mismos es por "negocio" y "solitario" por igual;

- Para Blue Soho Hotel, el segundo motivo de viaje más común entre los lusoparlantes es en “familia”; el tercer motivo es tanto con “amigos” y “solitario”; mientras que los viajes por negocio quedan al final.

De esta forma, fue posible constituir las siguientes hipótesis:

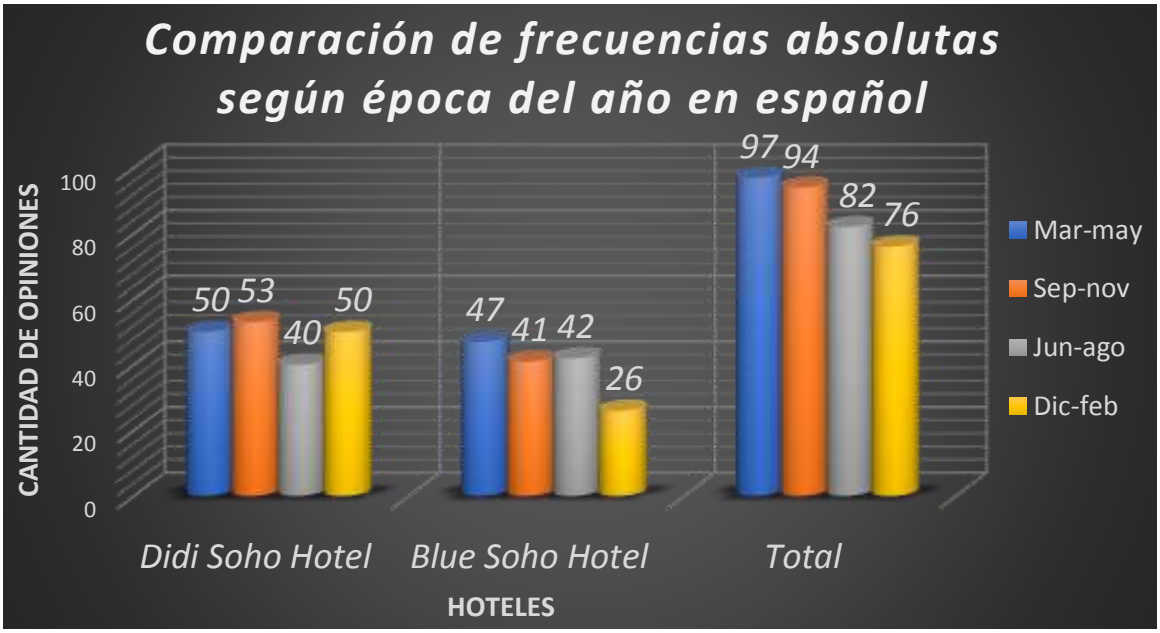
17. *“La gran mayoría de los turistas lusoparlantes que se hospedan en hoteles 3 estrellas de Palermo vienen en pareja”;*
18. *“El target principal de Blue Soho Hotel son los viajeros en pareja, sin importar la lengua en la que escriban los comentarios”;*
19. *“Para Didi Soho Hotel, el target principal de los viajeros lusoparlantes, son los viajeros con amigos”;*
20. *“Didi Soho Hotel posee un target de turistas menos definido que Blue Soho Hotel”;*
21. *“Los viajeros por negocios y solitarios, sin importar la lengua en la que escriban los comentarios, son los que menos se hospedan en hoteles 3 estrellas de Palermo”.*

5.1.3. Comparación de hoteles según la época del año.

5.1.3.1. En idioma español.

Mediante la utilización de los datos correspondientes a lo previamente explicado, se ha logrado confeccionar el siguiente gráfico:

Figura 1.10. Comparación de frecuencias absolutas según época del año en español



Fuente: elaboración propia en base a datos de TripAdvisor.

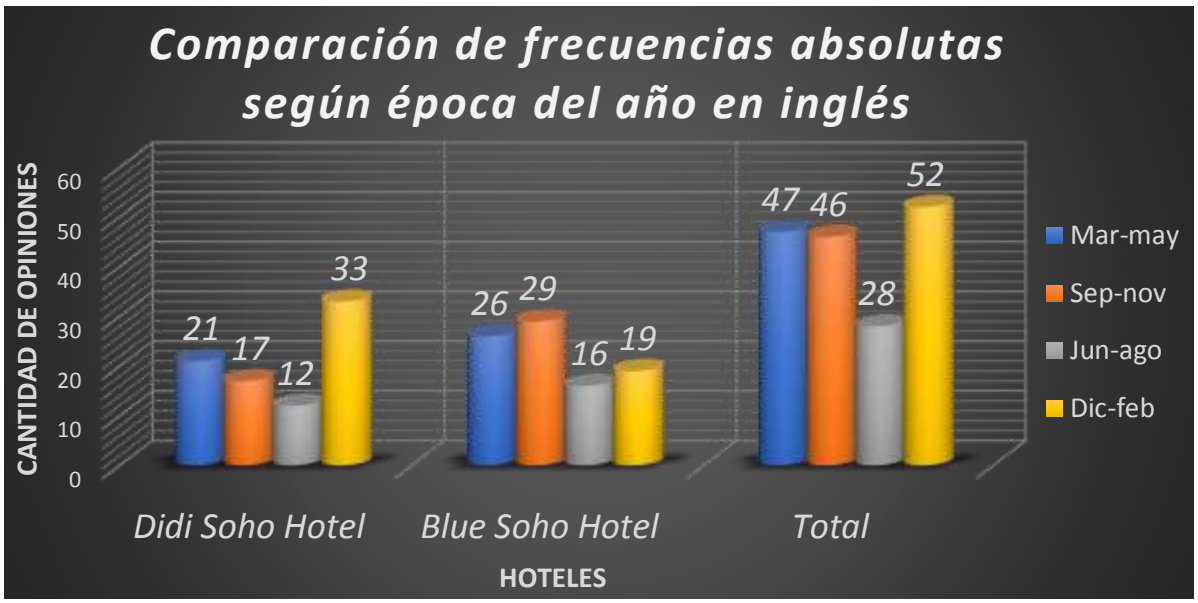
Visto el gráfico anterior, en donde se representa la cantidad de opiniones escritas en idioma español según la época del año para los dos hoteles, fue posible concluir que:

22. *"En Didi Soho Hotel se hospedan más hispanohablantes entre los meses de diciembre y febrero que en Blue Soho Hotel";*
23. *"En Blue Soho Hotel el período en donde menos se hospedan hispanohablantes es entre los meses de diciembre y febrero";*
24. *"No existe una época del año preferida para hospedarse en hoteles 3 estrellas de Palermo por los hispanohablantes".*

5.1.3.2. En idioma inglés.

Mediante la utilización de los datos correspondientes a lo previamente explicado, se ha logrado confeccionar el siguiente gráfico:

Figura 1.11. Comparación de frecuencias absolutas según época del año en inglés



Fuente: elaboración propia en base a datos de TripAdvisor.

Visto el gráfico anterior, en donde se representa la cantidad de opiniones escritas en idioma inglés según la época del año para los dos hoteles, fue posible concluir que:

25. *"En Didi Soho Hotel, el período en donde más angloparlantes se hospedan, es entre los meses de diciembre y febrero";*
26. *"En Blue Soho Hotel, entre los meses de diciembre y febrero, se hospedan menos angloparlantes que en Didi Soho Hotel";*

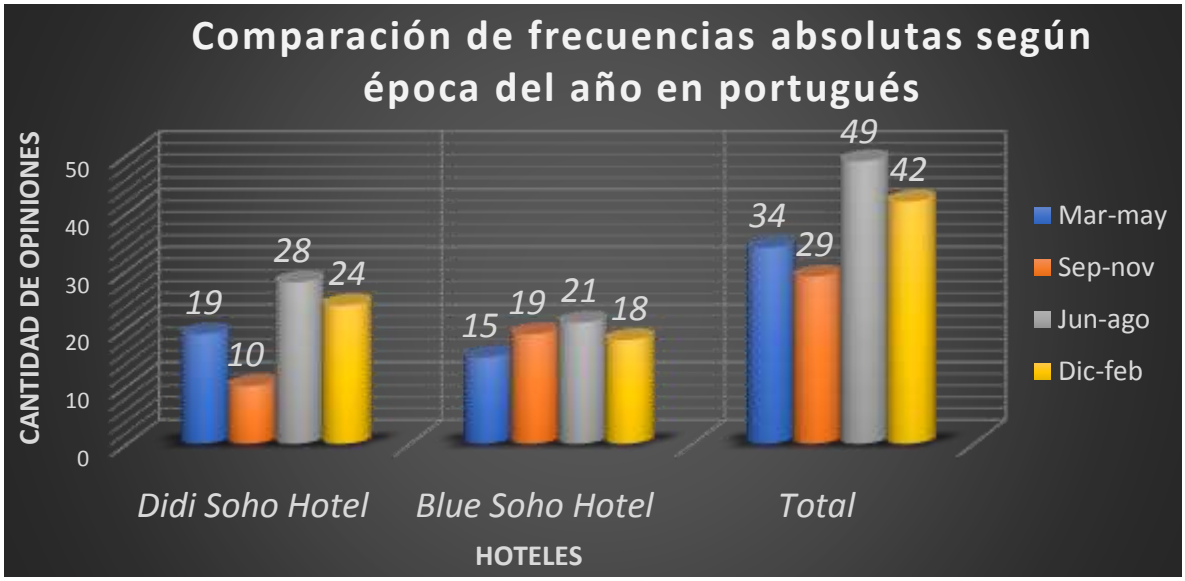
27. "En Blue Soho Hotel, entre los meses de septiembre y noviembre se hospedan más angloparlantes que en Didi Soho Hotel";

28. "Entre los meses de junio y agosto en cuando se hospedan menos turistas angloparlantes en hoteles 3 estrellas de Palermo".

5.1.3.3. En idioma portugués.

Mediante la utilización de los datos correspondientes a lo previamente explicado, se ha logrado confeccionar el siguiente gráfico:

Figura 1.12. Comparación de frecuencias absolutas según época del año en portugués



Fuente: elaboración propia en base a datos de TripAdvisor.

Visto el gráfico anterior, en donde se representa la cantidad de opiniones escritas en idioma portugués según la época del año para los dos hoteles, es posible concluir que:

29. "En Blue Soho Hotel, entre los meses de septiembre y noviembre se hospedan más lusoparlantes que en Didi Soho Hotel";

30. "Los lusoparlantes que se hospedan en hoteles 3 estrellas en Palermo no poseen una época del año preferida para viajar".

Así entonces se concluye con esta primera técnica que resume datos, muestra una rápida familiarización, y permite analizar y sacar conclusiones de estos datos.

5.2. Capítulo 2: Frecuencia de términos

5.2.1. Introducción y definición de pautas de trabajo.

Como ha explicado Witter et al. (2011) [i] la frecuencia de términos sirve para realizar un conteo de todos los términos utilizados en los documentos (conjuntos de opiniones, textos, etc.); que en este caso, correspondió la revisación de todas las opiniones. Además, muchos términos serán modificados con el propósito de facilitar el manejo de los datos, y de esta forma, se reduzcan ambigüedades y confusiones entre términos; y también, asegurar una correcta implementación de las herramientas y sus resultados.

Para ello se utilizó R[14] tanto para establecer cuántas veces apareció un término en la totalidad de las opiniones como para modificar los datos, y se tomó tanto el "texto" de la opinión como el "título" de la opinión para conformar un mismo documento, el cual fue sometido a estos experimentos y modificaciones.

Antes de proceder con la frecuencia de términos, se procedió con modificar los datos (como se explicó anteriormente) de acuerdo con las siguientes especificaciones para hacer más eficiente la tarea de frecuencia de términos:

- Se transformaron todas las letras mayúsculas a minúsculas;
- Se suprimieron todos los símbolos de puntuación;
- Se suprimieron todos los números no escritos con palabras;
- Se suprimieron todos los espacios que tengan un espacio antes;
- Se suprimieron las palabras de uso común para la lengua española[15];
- Se suprimieron otras palabras poco relevantes para el experimento[16].

Además, se sometieron las opiniones a otro tipo de modificaciones para que el procesamiento por parte de la misma técnica sea más eficiente. Principalmente, debido a la existencia del error

[14] Paquetes principales utilizados: "NLP", "tm", "wordcloud", "fpc", "cluster".

[15] Estas palabras vienen predeterminadas en un documento que ofrece el paquete "tm" de R.

[16] Ver listado completo de otros términos suprimidos en la sección de Anexos.

humano (Hohendahl, 2011)[17]. Por lo que, para evitar estas discrepancias a lo largo del trabajo se optó por modificar ciertas palabras para unificarlas un mismo término, y se lo hizo mediante tres formas[18]:

- Forma de la raíz: los plurales serán reducidos a singulares, verbos a sustantivos. Por ejemplo, la palabra "habitaciones" tomará la forma de "habitación", la palabra "cómoda", cómodas" y "cómodos" tomarán la forma de "cómodo", entre otros;
- Forma ortográfica: errores humanos de introducción de caracteres podrán ser atenuados. Por ejemplo, palabras como "habitacion", "habitacíon" y "habitacoin", tomarán la forma de "habitación", entre otros;
- Forma por similitud semántica: muchos sinónimos serán tratados bajo el mismo término. Por ejemplo "dormir" y "descansar" tomarán la forma de "descanso", "recepción", "staff" y "personal" tomarán la forma de "atención", entre otros.

Para las representaciones se optó por la utilización del "wordcloud", o literalmente 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; por otro lado, en cuanto al color, este se lo asigna según la frecuencia del término, por lo que el color[19] no agrupa palabras por su grado de relación.

5.2.2. Aplicación de la herramienta.

De esta forma, se procedió con realizar tanto la frecuencia de términos como las nubes de términos (los 50 términos más frecuentes)[20] correspondientes a Didi Soho Hotel y Blue Soho Hotel; y se obtuvieron los siguientes gráficos (en forma de nube de términos) también producidos con R:

[17] Como lo pueden ser los errores en el uso del teclado, no saber cómo se escribe una palabra, o también, por cuestiones de comodidad, puede que el usuario omita caracteres especiales como lo son las vocales con tilde y la letra "ñ".

[18] Ver listado completo de las modificaciones en la sección de Anexos.

[19] La escala de colores utilizada se llama "dark2" y ofrece los siguientes colores que se asignan de menor a mayor para las frecuencias: turquesa, naranja, violeta, fucsia, verde, amarillo, marrón, y gris.

[20] Ver tablas de frecuencias de términos en la sección de Anexos.

Figura 2.1. Nube de términos de Didi Soho Hotel.

Fuente: producción propia en base a datos de R

Para Didi Soho Hotel, los 10 términos más mencionados fueron: “bien” (con 144 repeticiones), “habitación” (125), “ubicación” (112), “hotel” (100), “atención” (78), “cómodo” (56), “excelente” (51), “grande” (46), “baño” (36), y “palermo” (35).

Figura 2.2. Nube de términos de Blue Soho Hotel.

Fuente: producción propia en base a datos de R

Para Blue Soho Hotel, los 10 términos más mencionados fueron: “bien” (con 139 repeticiones), “hotel” (130), “atención” (112), “habitación” (104), “ubicación” (82), “amable” (67), “muy” (50), “desayuno” (45), “excelente” (45), y “soho” (44).

5.2.3. Observaciones y análisis.

Para ambos hoteles, el término "bien" fue la más repetida, 144 veces para Didi Soho Hotel y 139 veces para Blue Soho Hotel; lo que no quiere decir necesariamente que Didi Soho Hotel posea este atributo en mayor intensidad que Blue Soho Hotel. Pues, al tratarse de datos y valores cualitativos no se puede asumir que una mayor cantidad de repeticiones equivalga a una mayor identificación con el término, es decir, que hay que asumir un determinado nivel de riesgo y/o incertidumbre.

Los siguientes términos que encabezan la lista (además de "bien") en ambos hoteles son: "hotel", "ubicación", "habitación", y "atención", aunque no necesariamente en este orden; pero sí determinan, visto a que se tratan de aspectos genéricos de un producto hotelero, sobre qué o de qué se habló con mayor recurrencia.

Por otro lado, se pueden identificar ciertos términos, los cuales son además adjetivos calificativos, estos son: "cómodo" y "grande" para Didi Soho Hotel, y "amable" para Blue Soho Hotel. Intuitivamente, y por relación semántica, los términos "cómodo" y "grande" estarían más ligados al término "habitación", y el término "amable" estaría más ligado al término "atención"; así, entonces, se podría intuir a priori que:

31. *"La ventaja competitiva de Didi Soho Hotel son las habitaciones" y;*
32. *"La ventaja competitiva de Blue Soho Hotel es la atención a los clientes por parte del personal".*

Sin embargo, aun asumiendo esto, el presente trabajo ofrece más adelante un estudio a mayor profundidad acerca de la relación de términos, y será presentado en el capítulo 3: "Asociación de términos", ya que se conoce si los adjetivos están empleados de tal forma que exprese una connotación negativa.

Siguiendo con los adjetivos, se puede identificar el término "excelente", el cual puede ser utilizado en un sentido muy amplio, pero semánticamente hablando no podría asumirse que su utilización está ligada a una palabra específica.

Otros de los términos más utilizados (pero en menor magnitud que las anteriores) son "baño" para Didi Soho Hotel, y "desayuno" para Blue Soho Hotel. Aquí, al no tener ningún otro término como guía a simple vista, no se podría asumir que su mención signifique un punto a

favor o un punto en contra de los respectivos hoteles, por lo que, visto esto, más adelante en el capítulo 3: Asociación de términos se podrá apreciar este interrogante más a fondo.

Por último, el término “muy” para Blue Soho Hotel está entre las 10 mayores frecuencias, siendo que dicho término en los datos recopilados representa la intensidad de un adjetivo o verbo, por lo que se podría asumir que el usuario manifestó desde su subjetividad que el estado de un algo o la ocurrencia de un hecho sucedió en niveles por encima o por debajo de lo que el mismo usuario consideraría como normal. Esta particularidad será estudiada a medida que se avance con la presentación de las distintas técnicas, precisamente en el capítulo 3: clasificación por análisis del sentimiento.

5.3. Capítulo 3: Asociación de términos

5.3.1. Introducción y criterios de procesamiento.

En este caso, se optó en lo posible por darle una total importancia a la asociación de términos del tipo semántico, que, según lo previsto, esta no requiere que cuestiones gramaticales sean tenidas en cuenta, sino que simplemente la relación semántica ocurre cuando un par de palabras ocurren en un mismo contexto (Pecina, 2009).

Como la asociación de términos otorga un valor a la coocurrencia entre dos términos, se previó que si se aplicara la técnica sobre las opiniones, un gran número de términos coocurrirían muy frecuentemente, pero no indicarían relación alguna, como por ejemplo, puede pasar que los usuarios digan que "la ubicación es excelente" y también que "la atención es mala", por lo tanto habría un índice de coocurrencia bastante alto entre "excelente" y "atención", lo cual sugiere un concepto erróneo acerca del hotel en cuestión. Por este motivo, para evitar asociaciones de términos semánticamente poco relacionados, se optará por, como señalan los autores citados, separar las opiniones en oraciones y suboraciones para someter a los datos a la asociación de términos (Pang y Lee, 2008, p. 29)[i],.

Para ello, se procedió con llamar al archivo .csv contenedor de las opiniones del hotel Didi Soho Hotel, y luego se separaron los conjuntos de cadenas de caracteres por el signo de puntuación "." (punto), y también por el signo de exclamación "!" simultáneamente. De esta forma, se logró separar las opiniones en oraciones simples y/o compuestas. Para esta tarea se utilizó R[21].

Luego de haber separado las opiniones en oraciones, se procedió con realizar los mismos seis tipos de modificaciones en las opiniones y tres tipos de modificaciones en las palabras (transformaciones, modificaciones y supresiones) realizadas en la aplicación de frecuencia de términos. Este paso en particular no se lo realizó al principio ya que esto evitaría que se pudieran separar las opiniones en oraciones, que ya los signos de puntuación se suprimirían con anterioridad.

[21] Principales paquetes utilizados: "NLP", "tm", "fpc", "cluster", "corrplot".

Finalmente, una vez obtenidos los índices de coocurrencia para los distintos términos, se los representará mediante gráficos de nodos[22] y dendrogramas.

5.3.2. Asociaciones con términos determinados.

En esta parte del trabajo se intentó localizar aquellos términos mayormente asociados con los principales aspectos los cuales un huésped común suele percibir sobre un producto hotelero, al igual que el nivel de asociación entre estos términos.

Para ello se tomaron los temas los cuales TripAdvisor menciona para ser calificados individualmente ("ubicación", "habitación", "atención", "limpieza", "calidad de descanso", y "relación calidad-precio"), y a estos se sumaron otros dos los cuales se consideraron importantes para ser explorados debido a su cantidad de menciones por los huéspedes ("mantenimiento" y "baño"). Por lo tanto, los aspectos establecidos a ser inspeccionados mediante esta técnica fueron en un total de 8 (ocho).

5.3.2.1. Asociaciones con el término "ubicación".

Con respecto al término "ubicación", se estableció que los demás términos[23] deben poseer una frecuencia mínima de 7 (siete) repeticiones, con un índice de coocurrencia con el término "ubicación" del 7,5%. De esta forma, los 5 primeros términos que el lenguaje R arroja, son los siguientes:

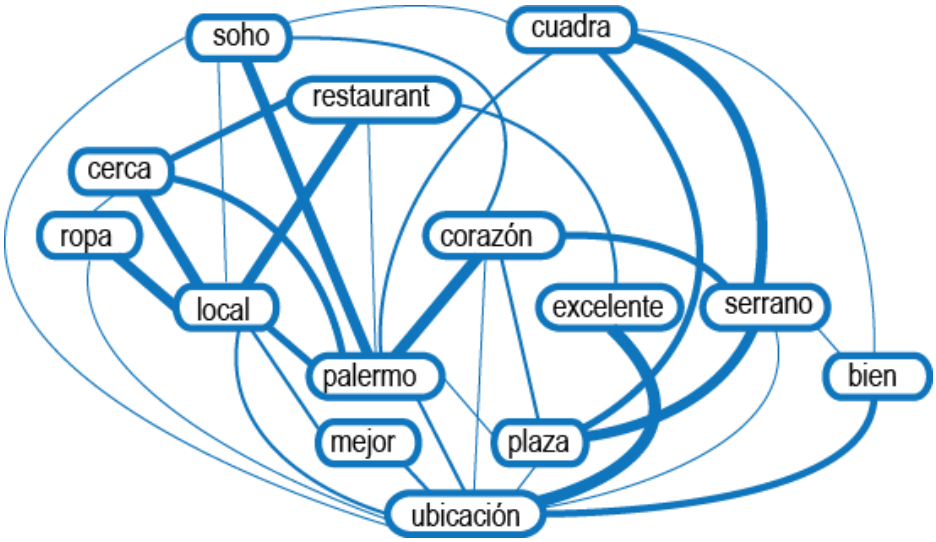
1. "excelente" – 47%;
2. "bien" – 32%;
3. "mejor" – 17%;
4. "palermo" – 17%;
5. "local" – 15%.

[22] Los gráficos de nodos fueron modificados con Adobe Illustrator ya que pueden presentar problemas en la impresión de un caracter especial, debido a que por el tipo de codificación que se utiliza está prevista para procesar únicamente el idioma inglés. Aun así, esto no representó ningún obstáculo para la visualización y la interpretación del gráfico, ya que su función más importante es mostrar el grado de relación entre dos términos, que lo hace con el grosor de las líneas.

[23] Ver todos los términos asociados con el término "ubicación" en la sección Anexos.

Teniendo en cuenta las especificaciones anteriores, y sumando que las asociaciones entre términos tengan un índice de coocurrencia mínima del 10%, se logró confeccionar el siguiente gráfico de nodos:

Figura 3.1. Asociación de términos con "ubicación"



Fuente: producción propia en base a datos de R

Visto la calidad semántica de cada término (que engloba a su vez a otros términos), la imagen producida y de acuerdo con las especificaciones sobre su interpretación, el investigador se permite inferir que:

1. La ubicación del hotel, al verse fuertemente relacionado con los términos "excelente", "bien" y "mejor", y de acuerdo con una escala ordinal decreciente por tratarse de datos cualitativos, puede identificarse con las siguientes proposiciones:

33. "La ubicación del hotel es excelente";
34. "La ubicación del hotel es buena", o bien, "el hotel está bien ubicado";
35. "El hotel tiene la mejor ubicación".

2. Además de términos que señalen una característica como "excelente", "bien" y "mejor", se puede detectar un cuarto término que es "cerca". Esta última está fuertemente asociada con los términos "restaurant", "local" (a su vez asociado con "ropa") y "palermo" (a su vez asociado con "mejor"). Visto esto, la imagen sugiere que:

36. "El hotel está cerca de restaurantes y locales de ropa";
37. "El hotel está en la mejor parte de Palermo".

3. Los términos “plaza” y “serrano” están fuertemente asociadas, pero según lo previsto solo se trata de una asociación por colocación (Pecina, 2009)[i], formando la popularmente conocida “Plaza Serrano”. Sin embargo, también están a su vez fuertemente relacionadas con el término “cuadra”. Por lo que esto sugiere que:

38. “El hotel está ubicado en, a una o más cuadras de Plaza Serrano”.

4. Los términos “palermo” y “soho” también poseen una asociación por colocación, formando el popularmente conocido Palermo Soho. A su vez, están fuertemente asociados con el término “corazón”. Por lo que esto sugiere que:

39. “El hotel está ubicado en el corazón de Palermo Soho”.

5.3.2.2. Asociaciones con el término “habitación”.

Con respecto al término “habitación”, se estableció que los demás términos[24] deben poseer una frecuencia mínima de 12 (doce) repeticiones, con un índice de coocurrencia con el término “habitación” del 15%. De esta forma, los 5 primeros términos que el lenguaje R arroja, son los siguientes:

1. “grande” – 43%;
2. “cómodo” –27%;
3. “cama” – 19%;
4. “chico” – 18%;
5. “baño” – 17%.

Teniendo en cuenta las especificaciones anteriores, y sumando que las asociaciones entre términos tengan un índice de coocurrencia mínima del 7,5%, y obviando términos poco relevantes como: “hotel” y “amable”, se logró confeccionar el siguiente gráfico de nodos:

[24] Ver todos los términos asociados con el término “habitación” en la sección Anexos.

Figura 3.2. Asociación de términos con "habitación"

Fuente: producción propia en base a datos de R

Visto la calidad semántica de cada término (que engloba a su vez a otros términos), la imagen producida y de acuerdo con las especificaciones sobre su interpretación, el investigador se permite inferir que:

1. La habitación o las habitaciones se ven fuertemente relacionados con los términos "grande", "cómodo" y "lindo" (a su vez relacionado con el término "muy") y de acuerdo con una escala ordinal decreciente, debido a que contamos con datos cualitativos, son apreciadas de la siguiente forma:

40. "las habitaciones son grandes";
41. "las habitaciones son cómodas";
42. "las habitaciones son lindas";
43. "las habitaciones son muy lindas".

2. Viendo la posible relación entre "grande" y "cómodo", ambos sugieren que cuando un usuario se refiere a la calidad de la habitación como "grande", es probable que también se refiere a esta como "cómoda", es decir, pueden suceder simultáneamente. Por lo tanto, es posible construir la siguiente proposición:

44. "Las habitaciones son grandes y cómodas".

3. 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 un lado, sería contradictorio que un huésped se refiera a la habitación como "grande" y "chico", por otro lado, también sería contradictorio que un huésped se refiera al baño como "grande" y "chico". Por lo tanto, se concluye lo siguiente:

45. Con un gran nivel de confianza: "la habitación es grande, pero el baño es chico";

46. Con un nivel de confianza menor: "la habitación es chica, y el año es grande".

4. Además, la habitación o las habitaciones presentan una fuerte asociación con los términos "desayuno" y "ruido". Estos términos serán tratados más adelante; pero ambos podrían señalar lo siguiente:

47. "El desayuno se sirve en la habitación";

48. y "Desde las habitaciones se pueden escuchar ruidos".

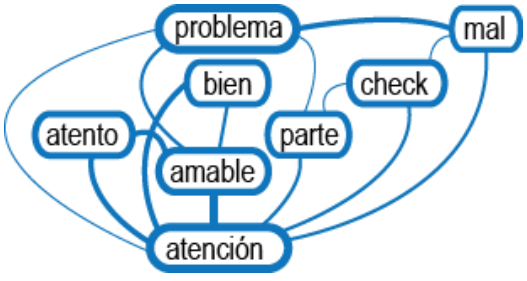
5.3.2.3. Asociaciones con el término "atención".

Con respecto al término "atención", se estableció que los demás términos[25] deben poseer una frecuencia mínima de 5 (cinco) repeticiones, con un índice de coocurrencia con el término "atención" del 12%. De esta forma, los 5 primeros términos que el lenguaje R arroja son los siguientes:

1. "amable" – 31%;
2. "parte" –26%;
3. "atento" – 23%;
4. "bien" – 21%;
5. "check" – 16%.

Teniendo en cuenta las especificaciones anteriores, y sumando que las asociaciones entre términos tengan un índice de coocurrencia mínima del 7,5%, se logró confeccionar el siguiente gráfico de nodos:

Figura 3.3. Asociación de términos con "atención"



Fuente: producción propia en base a datos de R

[25] Ver todos los términos asociados con el término "atención" en la sección Anexos.

Visto la calidad semántica de cada término (que engloba a su vez a otros términos), la imagen producida y de acuerdo con las especificaciones sobre su interpretación, el investigador se permite inferir que:

1. La atención por parte del personal (incluye dueños y/o gerentes), al estar fuertemente relacionado con los términos "amable", "atento", "parte" y "bien" y de acuerdo con una escala ordinal decreciente, debido a que contamos con datos cualitativos, son apreciadas de la siguiente forma:

49. "El personal es amable";
50. "El personal es atento";
51. "Hay buena atención por parte del personal".

2. Con respecto al término "parte", esta se puede interpretar de dos formas. La primera, es la prevista en la anterior, en donde por parte del personal hubo una buena atención; la segunda, visto que también está relacionado por el término "problema" y "check", se puede concluir lo siguiente:

52. "El servicio fue bueno en parte, salvo por un problema, que tuvo que ver con el check-in/out".

5.3.2.4. Asociaciones con el término "limpieza".

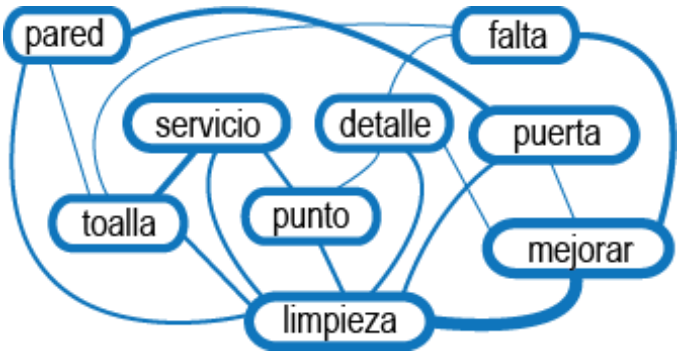
Con respecto al término "limpieza", se estableció que los demás términos[26] deben poseer una frecuencia mínima de 5 (cinco) repeticiones, con un índice de coocurrencia con el término "limpieza" del 11%. De esta forma, los 5 primeros términos que el lenguaje R arroja son los siguientes:

1. "mejorar" – 41%;
2. "puerta" –24%;
3. "toalla" – 24%;
4. "punto" – 21%;
5. "detalle" – 17%.

[26] Ver todos los términos asociados con el término "limpieza" en la sección Anexos.

Teniendo en cuenta las especificaciones anteriores, y sumando que las asociaciones entre términos tengan un índice de coocurrencia mínima del 7,5%, y obviando términos poco relevantes como: “aunque”, “atención” y “amable”, se logró confeccionar el siguiente gráfico de nodos:

Figura 3.4. Asociación de términos con “limpieza”



Fuente: producción propia en base a datos de R

Visto la calidad semántica de cada término (que engloba a su vez a otros términos), la imagen producida y de acuerdo con las especificaciones sobre su interpretación, el investigador se permite inferir que:

1. La relación “limpieza” y “mejorar”, nos dice claramente lo siguiente:

   53. “La limpieza debe mejorar”.

2. El término “falta” está íntimamente relacionado con el término “mejorar”, en donde se asume que:

   54. “Falta mejorar la limpieza”.

3. Los términos “punto” y “detalle”, simplemente sirven para señalar que podría existir algún inconveniente que el huésped haya detectado, por lo que su discurso los emplearía de las siguientes formas:

   55. “Hay un punto con respecto a la limpieza”;
   56. “Existe un detalle sobre la limpieza”;

4. Otros términos como “toalla”, “puerta” y “pared”, señalan que posiblemente el problema con la limpieza esté relacionado con estos objetos, por lo que:

   57. “los problemas en la limpieza están en la toalla, puerta y pared”.

5.3.2.5. Asociaciones con el término "mantenimiento".

Con respecto al término "mantenimiento", se estableció que los demás términos[27] deben poseer una frecuencia mínima de 5 (cinco) repeticiones, con un índice de coocurrencia con el término "mantenimiento" del 6%. De esta forma, los 5 primeros términos que el lenguaje R arroja son los siguientes:

1. "falta" – 29%;
2. "detalle" –25%;
3. "mal" – 17%;
4. "baño" – 14%;
5. "mejorar" – 11%.

Teniendo en cuenta las especificaciones anteriores, y sumando que las asociaciones entre términos tengan un índice de coocurrencia mínima del 12,5%, y obviando términos poco relevantes como: "amable", se logró confeccionar el siguiente gráfico de nodos:

Figura 3.5. Asociación de términos con "mantenimiento"

Fuente: producción propia en base a datos de R

Visto la calidad semántica de cada término (que engloba a su vez a otros términos), la imagen producida y de acuerdo con las especificaciones sobre su interpretación, el investigador se permite inferir que:

[27] Ver todos los términos asociados con el término "mantenimiento" en la sección Anexos.

1. El mantenimiento está íntimamente relacionado con los términos "falta" (a su vez relacionado con los términos "mejorar" y "limpieza") y "mal", del cual podemos concluir que:

58. "El hotel está mal mantenido";

59. y "como falta mejorar la limpieza, esto repercute en el nivel de mantenimiento".

2. Con respecto al término "detalle", a su vez relacionado con los términos "baño", la cual a su vez también está relacionado con los términos "pared", "puerta" y "ducha", se podría asumir que:

60. "El nivel de mantenimiento está condicionado con uno o más detalles en el baño de las habitaciones, estos detalles yacen tanto en la pared, en la puerta y en la ducha".

5.3.2.6. Asociaciones con el término "precio".

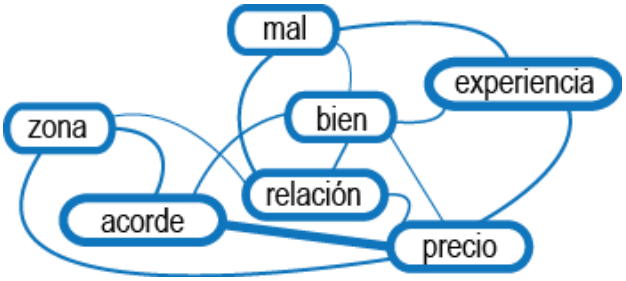
Con respecto al término "precio", se estableció que los demás términos[28] deben poseer una frecuencia mínima de 4 (cuatro) repeticiones, con un índice de coocurrencia con el término "precio" del 5%. De esta forma, los 5 primeros términos que el lenguaje R arroja son los siguientes:

1. "acorde" – 30%;
2. "relación" –20%;
3. "zona" – 18%;
4. "bien" – 14%;
5. "experiencia" – 10%.

Teniendo en cuenta las especificaciones anteriores, y sumando que las asociaciones entre términos tengan un índice de coocurrencia mínima del 5%, y obviando términos poco relevantes como: "pareció", "cualquier", y "pasar", se logró confeccionar el siguiente gráfico de nodos:

[28] Ver todos los términos asociados con el término "precio" en la sección Anexos.

Figura 3.6. Asociación de términos con "precio"



Fuente: producción propia en base a datos de R

Visto la calidad semántica de cada término (que engloba a su vez a otros términos), la imagen producida y de acuerdo con las especificaciones sobre su interpretación, el investigador se permite inferir que:

1. Los adjetivos calificativos relacionados con el término "precio" son: "acorde", "bien" y "mal". Por lo que podemos asumir, en función de una escala ordinal decreciente, las siguientes 3 proposiciones:

   61. "El precio por una habitación es acorde";
   62. "El precio por una habitación es bueno";
   63. "El precio por una habitación es malo".

2. Por otro lado, vemos que en el caso del término "acorde", este está relacionado con los términos "zona" y "bien", y a su vez el término "zona" está fuertemente relacionado con el término "precio". Por lo tanto, podemos asumir la siguiente proposición:

   64. "El precio por una habitación es acorde debido a la zona".

3. Por otro lado, con respecto a las asociaciones entre los términos "precio" con "relación", "mal", "bien" y "experiencia", se podría asumir la siguiente proposición:

   65. "A pesar de una mala experiencia, la relación precio calidad es buena".

5.3.2.7. Asociaciones con el término "descansar".

Con respecto al término "descansar", se estableció que los demás términos[29] deben poseer una frecuencia mínima de 9 (nueve) repeticiones, con un índice de coocurrencia con el término

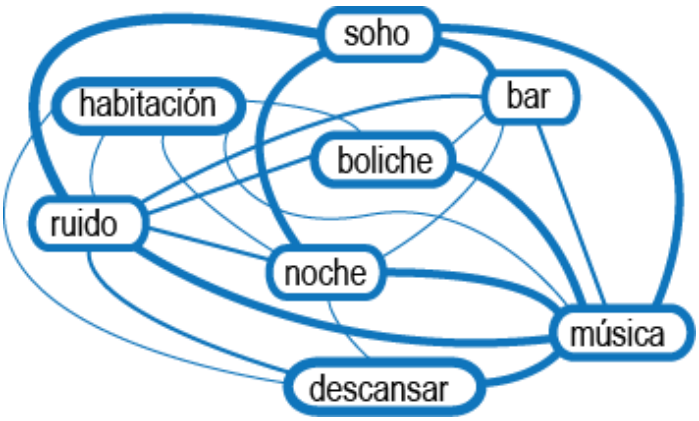
[29] Ver todos los términos asociados con el término "descansar" en la sección Anexos.

"descansar" del 5%. De esta forma, los 5 primeros términos que el lenguaje R arroja son los siguientes:

1. "música" – 28%;
2. "noche" –20%;
3. "ruido" – 17%;
4. "boliche" – 10%;
5. "habitación" – 10%.

Teniendo en cuenta las especificaciones anteriores, y sumando que las asociaciones entre términos tengan un índice de coocurrencia mínima del 7,5%, y obviando términos poco relevantes como: "limpio", "pava", y "eléctrica", se logró confeccionar el siguiente gráfico de nodos:

Figura 3.7. Asociación de términos con "descansar"



Fuente: producción propia en base a datos de R

Visto la calidad semántica de cada término (que engloba a su vez a otros términos), la imagen producida y de acuerdo con las especificaciones sobre su interpretación el investigador se permite inferir que:

1. Aquellos términos directa y fuertemente relacionados con el término "descansar" son: "música", "noche" y "ruido". Los cuales también están fuertemente relacionados con los términos "boliche", "bar", "lado" y "habitación". De esta forma, al tener todos los términos con asociaciones tan evidentes, podemos asumir las siguientes proposiciones:

66. *"La calidad de descanso se ve afectada por las noches, debido al ruido y a la música que se puede escuchar desde la habitación, en donde al parecer, se encuentra al lado de un bar o boliche".*

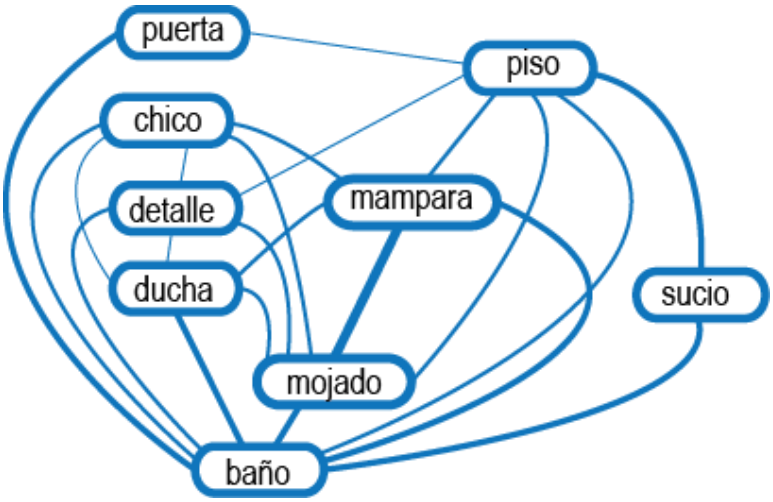
5.3.2.8. Asociaciones con el término "baño".

Con respecto al término "baño", se estableció que los demás términos[30] deben poseer una frecuencia mínima de 4 (cuatro) repeticiones, con un índice de coocurrencia con el término "baño" del 16%. De esta forma, los 5 primeros términos que el lenguaje R arroja son los siguientes:

1. "mojado" – 46%;
2. "ducha" –38%;
3. "detalle" – 35%;
4. "mampara" – 32%;
5. "sucio" – 25%.

Teniendo en cuenta las especificaciones anteriores, y sumando que las asociaciones entre términos tengan un índice de coocurrencia mínima del 5%, y obviando términos poco relevantes como: "foto", "habitación" y "departamento", se logró confeccionar el siguiente gráfico de nodos:

Figura 3.8. Asociación de términos con "baño"



Fuente: producción propia en base a datos de R

Visto la calidad semántica de cada término (que engloba a su vez a otros términos), la imagen producida y de acuerdo con las especificaciones sobre su interpretación, el investigador se permite inferir que:

[30] Ver todos los términos asociados con el término "baño" en la sección Anexos.

1. Los adjetivos calificativos más relacionados con el baño son: “mojado”, “sucio” y “chico”. Esto supone las siguientes proposiciones:

67. “El baño por alguna razón está mojado o se moja fácilmente”;
68. “El baño es sucio o posee suciedad”;
69. “El baño es chico”.

2. Con respecto al término “mojado”, está también está relacionado con los términos “mampara”, “ducha”, “piso”, “detalle”, y “chico”, y estos términos también se relacionan entre ellos. Por lo que, debido a que probablemente pertenezcan a una misma proposición, se pueden construir las siguientes:

70. “El baño queda mojado debido a la mampara”;
71. “Hay un detalle con respecto a la ducha, el cual moja el piso”;
72. “Debido a que el baño es chico, es fácil que la ducha moje el baño”.

3. Con respecto al término “puerta”, esta se relaciona solo con los términos “baño” y “piso”. Por lo que se puede asumir que:

73. “Existe un detalle con respecto a la puerta del baño”;
74. O bien, “Debido a que el baño es chico, la puerta obstruye el camino y/o es representa un inconveniente”.

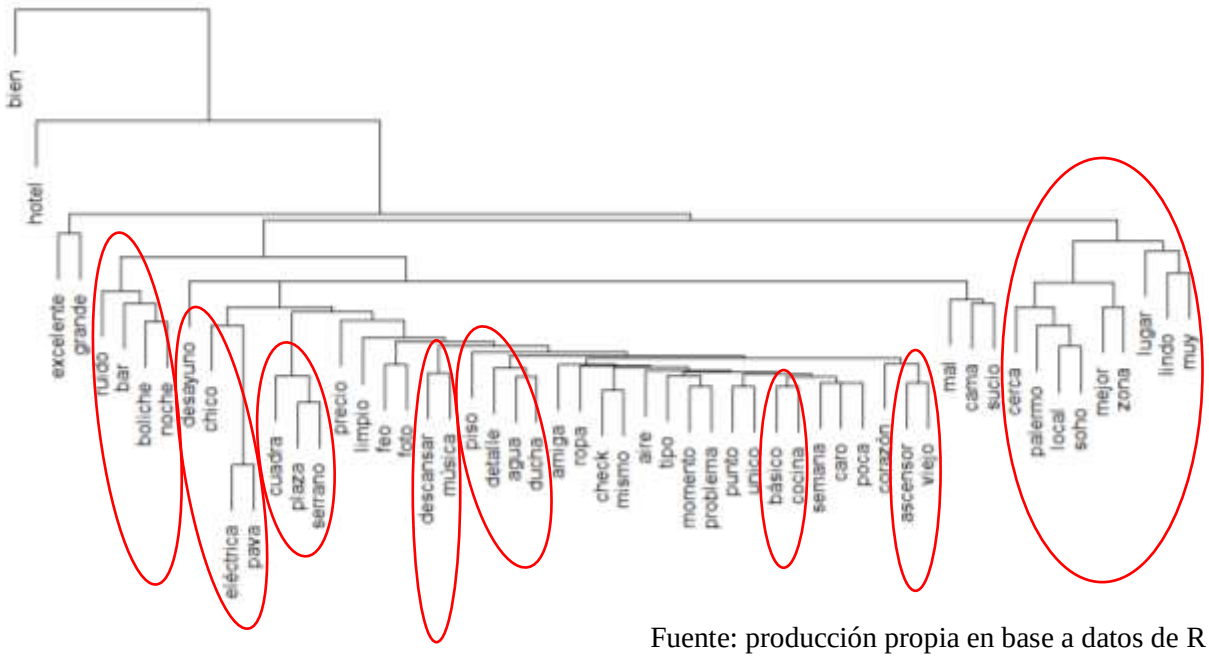
5.3.3. Asociaciones con términos indeterminados.

Para esta parte, se intentó crear gráficos que representen la asociación entre términos, sin especificar con cuáles términos deben estar asociados, por lo que, automáticamente, el lenguaje R otorga prioridad a aquellos términos que están más fuertemente asociados y menos distanciados.

De esta forma, mediante un gráfico de dendrogramas[31] y limitando la herramienta a solamente aquellos términos que poseen una frecuencia mínima de 5 repeticiones, se logró imprimir la siguiente imagen:

[31] Diversas fuentes definen los dendrogramas como un diagrama de datos en forma de árbol que organiza los datos en subcategorías que se van dividiendo en otros hasta llegar a un nivel de detalle deseado. En esta se puede apreciar claramente el grado de relación entre los datos, e incluso entre

Figura 3.9. Asociación de términos no discriminados



Fuente: producción propia en base a datos de R

En esta imagen, obviando las asociaciones por colocación, se pueden evidenciar a simple vista un buen número de grupos de términos con fuertes asociaciones entre sus términos, y estos grupos son los siguientes:

1. “ruido”, “bar”, “boliche”, y “noche”;
2. “pava”, “eléctrica”, “desayuno”, y “chico”;
3. “plaza”, “serrano”, y “cuadra”;
4. “descansar”, y “música”;
5. “ducha”, “agua”, “detalle”, y “piso”;
6. “cocina”, y “básico”;
7. “ascensor”, y “viejo”;
8. “local”, “cerca”, “palermo”, “soho”, “mejor”, “zona”, “muy”, “lindo”, y “lugar”.

Algunos de estos grupos de términos son fácilmente identificables debido a que las combinaciones entre estos términos se asemejan a proposiciones hechas anteriormente en el punto de asociaciones de términos discriminados. De hecho, al realizarse el análisis de estos, se volvió a referirse a las proposiciones hechas en las asociaciones de términos con términos

---

grupos de datos habiendo o no similitud o cercanía entre categorías. Para esta técnica, se utilizó el algoritmo Euclidiano, en donde mientras más cerca estén dos términos y mientras menos ramas se abran de uno al otro, más relacionados están.

discriminados. Sin embargo, existen grupos de términos que no fueron identificados, o bien, no se han establecido de tal forma que sus términos tengan asociaciones claras y evidentes.

Una vez que se determinaron estos grupos, se procedió con realizar las siguientes observaciones, comparaciones y análisis:

1. Al igual que en el análisis propuesto en el punto de asociación de términos discriminando el término "descansar" (que se refiere a la calidad de descanso), se puede observar cómo estos cuatro términos ("ruido", "bar", "boliche", y "noche") sugieren y apoyan la proposición previamente elaborada, y a su vez, se pudo producir lo siguiente:

75. "hay ruidos durante las noches provenientes de bares y boliches que evitan que uno descanse";

2. En la asociación de términos discriminando el término "habitación", se hizo una breve mención en la posible asociación entre los términos "desayuno" y "chico". Que para este caso, visto la asociación entre los cuatro términos de este punto ("pava", "eléctrica", "desayuno", y "chico"), se observa que:

76. "el desayuno en las habitaciones es modesto e incluye una pava eléctrica";

3. Retomando la asociación de términos discriminando el término "ubicación", se vio una vez más que los tres términos de este punto ("plaza", "serrano", y "cuadra") dieron lugar a la construcción de la siguiente proposición:

77. "el hotel queda situado a una cuadra de Plaza Serrano";

4. Al igual que el primer punto, y refiriéndose a la asociación de términos discriminando el término "descansar", los dos términos de este cuarto punto ("descansar", y "música"), dieron lugar a lo siguiente[32]:

78. "es difícil descansar por la música";

[32] También no es difícil pensar en la siguiente proposición: *"la música permite descansar"*, pero por experiencias anteriores del investigador, por lo propuesto durante la confección de la presente investigación, y por medio del uso del sentido común, esta la proposición específica fue descartada.

5. Volviendo a la asociación de términos discriminando el término “baño”, y tomando los cuatro términos que corresponden a este punto (“ducha”, “agua”, “detalle”, y “piso”) se puede pensar en las siguientes proposiciones similares:

79. “hay un detalle con respecto a la ducha, en donde el agua deja el piso mojado”;

6. Para este sexto grupo, no fue posible identificar ninguna asociación entre estos dos términos (“cocina”, y “básico”) en puntos anteriores, por lo que quedó con proponer lo siguiente:

80. “la cocina es básica”;
81. o bien, “para cocinar uno tiene lo básico”;

7. Para este séptimo grupo, tampoco fue posible identificar ninguna asociación entre los dos términos (“ascensor”, y “viejo”), sin embargo, a diferencia del sexto grupo, realizar una proposición es relativamente fácil, y se obtiene la siguiente construcción:

82. “el ascensor es viejo o antiguo”;

8. En este último grupo, claramente el tema central fue la “ubicación”, en donde al igual que en el punto de asociación de términos discriminando el mismo término “ubicación”, se puede ver la similitud con las proposiciones anteriores; por lo que, se consideró mediante un razonamiento lógico que:

83. “el hotel está ubicado en un muy lindo lugar en Palermo Soho, en la mejor zona, donde uno está cerca de los locales”.

5.4. Capítulo 4: Clasificación por sentimiento

5.4.1. Introducción a la clasificación por sentimiento.

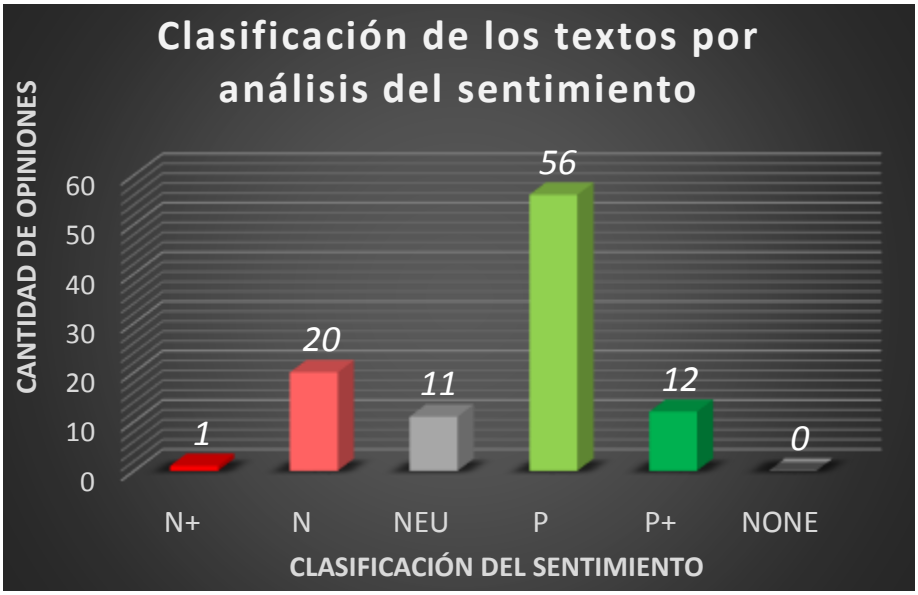
Como bien explica Bing (2012, p. 31), la clasificación por sentimiento consta en clasificar el texto en dos tipos principales de opiniones: opiniones positivas y opiniones negativas, y de acuerdo con las palabras que expresen un sentimiento u opinión. Mientras el vocabulario utilizado tienda a expresar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

Para poner en prueba esta técnica se utilizaron ambos programas: Excel (extensión MeaningCloud) y R, cada uno con sus propias funciones.

5.4.2. Mediante funciones proporcionadas por Excel.

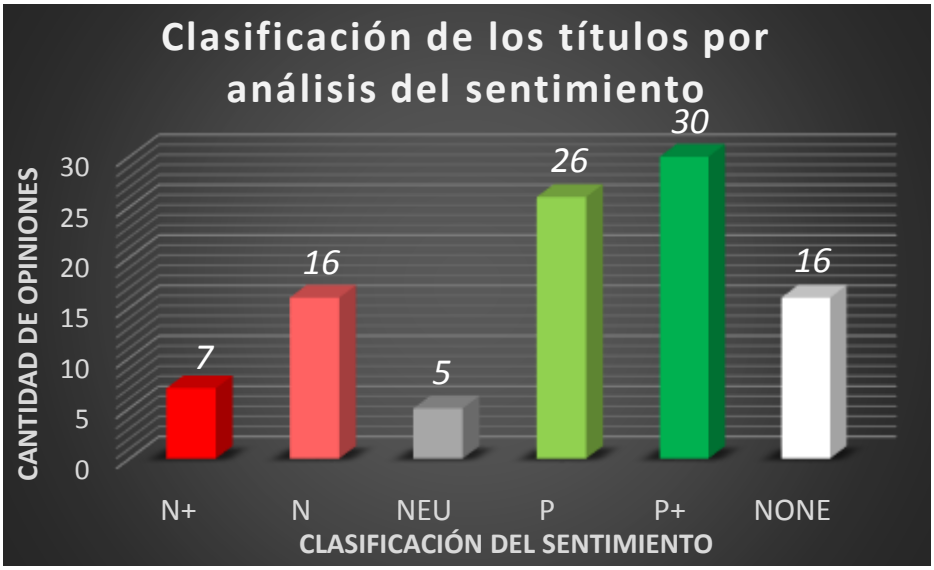
Para esta tarea de clasificación por sentimiento, se utilizó la extensión MeaningCloud de Excel. Esta herramienta procesa un campo del tipo texto y puede arrojar seis posibilidades diferentes, que en orden creciente de polaridad, las primeras cinco son: N+ (muy negativo), N (negativo), NEU (neutro), P (positivo), P+ (muy positivo), y la última se trata de otra categoría aparte que es NONE (no clasificable). Dicho esto, esta función de clasificación fue aplicada tanto para el título de la opinión como para el texto de la opinión. De esta forma, se confeccionaron los siguientes gráficos de acuerdo con los resultados obtenidos:

Figura 4.1. Clasificación de los textos por análisis del sentimiento



Fuente: producción propia en base a datos de Excel

Figura 4.2. Clasificación de los títulos por análisis del sentimiento



Fuente: producción propia en base a datos de Excel

Estos datos obtenidos resultaron no ser de utilidad, ya que no se pudieron detectar comportamientos similares entre estas dos variables (sentimiento por texto y sentimiento por título). Por este motivo, se procedió con intentar unificarlos en una sola variable.

Para determinar la clasificación por sentimiento para cada opinión, se tomaron los resultados del análisis del sentimiento para tanto el título de la opinión como el texto de la opinión, y también, el puntaje general otorgado por el usuario.

Se determinó que los resultados del análisis del sentimiento siguieran una escala del tipo numérica, por lo que se los transformó mediante la siguiente tabla:

Figura 4.3. Escala de transformación de la clasificación por análisis del sentimiento

	Valores					
Clasificación por análisis del sentimiento	N+	N	NEU	P	P+	NONE
Clasificación del huésped	1	2	3	4	5	No considerado

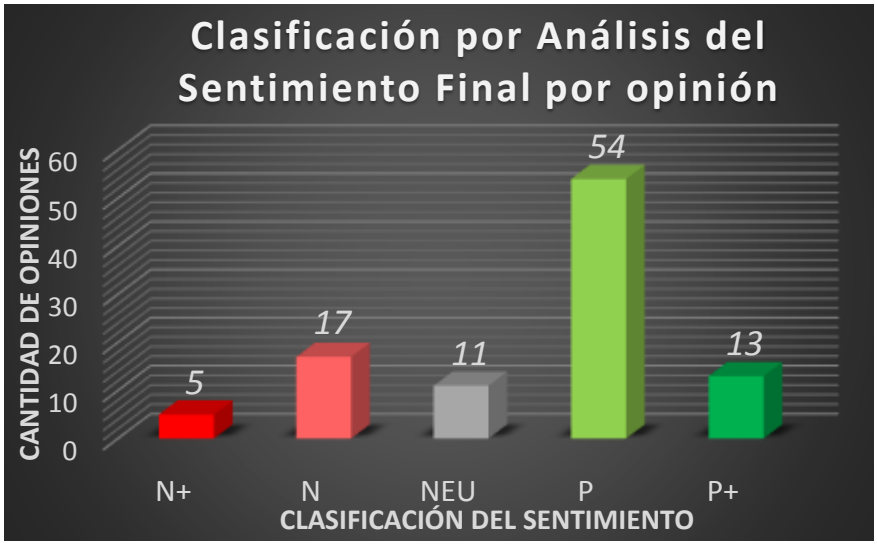
Fuente: escala para transformación numérica confeccionado por el investigador

Una vez que se obtuvieron los puntajes del título y del texto para cada una de las opiniones transformados a escala numérica, y los puntajes otorgados por cada usuario, se procedió con la utilización de diferentes medidas de tendencia central para determinar la clasificación por análisis del sentimiento final para cada opinión.

De esta forma, de tres valores para cada opinión, se determinó el puntaje mediante la Mo(x), dando así una mayor importancia a aquel valor que más aparece; y en caso de no poder determinarse mediante esta, se utilizó la Me(x), evitando que los extremos tanto superior como inferior afecten al valor final.

Una vez que se tiene el valor final, este se vuelve a transformar mediante la misma escala numérica con la cual fue transformada; y de esta forma se obtuvo el siguiente gráfico para representar la clasificación por análisis del sentimiento de las opiniones:

Figura 4.4. Clasificación por análisis del sentimiento final por opinión



Fuente: producción propia en base a datos de Excel

Debido a que la escala utilizada (N+; N; NEU; P; y P+) pertenece a una variable del tipo ordinal, se desconocen el valor de los intervalos entre cada orden, tampoco se pueden someter dos o más valores a operaciones aritméticas para obtener otro valor equivalente a un orden, y no es posible detectar un cero absoluto.

Como observación detectada que posiblemente podría tomarse como información es con respecto a la Mo(x) de la clasificación del sentimiento: P (con una frecuencia de 54) que ocupa la cuarta posición en cuanto al valor (suponiendo una escala ascendente de valores); y esta cuarta posición coincide con la posición de la Mo(x) detectada en el capítulo 1 (puntajes de los turistas en idioma español), en donde la Mo(x) muy bueno. Para este caso podría considerarse que, ambos P y "muy bueno" al ser las modas de una variable de 5 valores y ocupando la 4ta posición (suponiendo que ambas variables están ordenadas de forma ascendente), pueden referirse a la mayoría los turistas y pueden ser representativas de la población de turistas para

Didi Soho Hotel. De hecho, la muestra sostiene que 30 de las 34 opiniones calificadas como "muy bueno" pertenecen al grupo de las 54 clasificadas como P.

Por otro lado, también podría suponerse una similitud entre las poblaciones en cuanto a la "calificación de los usuarios" y la "clasificación por análisis del sentimiento de los usuarios". De esta forma, ambos tendrían una μ, Me(x) y Mo(x) similares, y también tendrían un coeficiente de asimetría negativo (y por lo tanto asimétrica hacia la izquierda) y un coeficiente de curtosis negativo (poco empinamiento o platicúrtico). Por lo que podría proponerse, en líneas generales, que:

84. "A pesar de los puntajes y aspectos negativos, los huéspedes por lo general están satisfechos con el producto hotelero que ofrece Didi Soho Hotel".

5.4.2. Mediante funciones proporcionadas por R.

Para esta tarea, en donde se procedió con escribir un código a priori[33], muy poco potente a comparación de otras herramientas de análisis del sentimiento, pero lo suficiente para realizar el análisis del sentimiento mediante R[34], lo que se hizo fue asignar un término extra a los términos que señalan una connotación positiva y negativa. Para el caso de los términos positivos, el término extra asignado es "positivo" y para el caso de los términos negativos, el término extra asignado fue "negativo". Además, se agregó el término "potenciado", que refleja la potencia con la cual el usuario intensifica la calidad semántica de su discurso.[35]

Se procedió entonces con establecer los índices de coocurrencia con los términos "positivos" y "negativo", y se lo hizo con los términos los cuales se consideraron interesantes para conocer la polaridad del sentimiento del usuario, estos, todos sustantivos cuyas frecuencias de ocurrencia fueron las más altas (ver "capítulo 2: Frecuencia de términos"), son, en orden alfabético: "ascensor", "atención", "baño", "bar", "boliche", "café", "cama", "cocina", "desayuno", "descansar", "detalle", "ducha", "habitación", "hotel", "limpieza",

[33] Uno de los mayores problemas para la investigación fue que aunque la técnica y sus funciones de análisis del sentimiento estén disponibles para el lenguaje R, esta funciona siempre y cuando el lenguaje natural a procesar sea el inglés.

[34] Principales paquetes utilizados: "NLP", "tm", "rlist" y "corrplot".

[35] Ver términos que tuvieron una asignación de términos extra en la sección Anexos.

"mantenimiento", "música", "noche", "piso", "precio", "problema", "relación", "ruido", y "ubicación".[36]

Para la representación y visualización de los resultados se utilizó el gráfico de matriz de correlación mediante R[37], en donde la intensidad del color azul representa un índice de coocurrencia significativo, y la intensidad del color rojo representa un índice inverso de coocurrencia significativo. Esta imagen básicamente refleja que tan "negativa" o "positiva" en cuanto al sentimiento es la utilización de un término.

Figura 4.5. Clasificación por análisis del sentimiento por términos determinados

Fuente: producción propia en base a datos de R

Teniendo esta imagen creada y los términos que presenten asociaciones considerables con los términos "negativo" y "positivo", se pudo deducir lo siguiente con respecto al análisis del sentimiento mediante la herramienta R:

- Los términos que más evidenciaron un ajuste hacia una polaridad negativa son, en orden descendente, fueron: "ruido", "música", "baño", "problema", "habitación", "hotel", y "descansar".
- Además, los términos que menos evidenciaron ajustarse hacia una polaridad negativa, es decir, una asociación negativa con lo "negativo", fueron: "ubicación", y "cocina".
- Por otro lado, los términos que más evidenciaron ajustarse hacia una polaridad positiva, en orden descendente, fueron: habitación", "potenciado", "ubicación", "cama", y "atención".

[36] Ver los índices de coocurrencia en la sección Anexos.

[37] Aunque el nombre de la herramienta de visualización mencione la correlación en él, también puede ser utilizada para representar las coocurrencias.

- También, aquellos términos que menos evidenciaron ajustarse hacia una polaridad positiva, es decir, una asociación negativa con lo "positivo", fueron: "ascensor", "música", "piso", y "ruido".

Visto esto, se puede constituir lo siguiente en materia de proposiciones e información generada:

85. "La utilización del término "ruido" es lo que más afecta negativamente a las calificaciones de Didi Soho Hotel";

86. "La utilización del término "música" afecta negativamente a las calificaciones de Didi Soho Hotel";

87. "La utilización del término "baño" afecta negativamente a las calificaciones de Didi Soho Hotel";

88. "La utilización del término "problema" afecta negativamente a las calificaciones de Didi Soho Hotel";

89. "La utilización del término "hotel" afecta negativamente a las calificaciones de Didi Soho Hotel";

90. "La utilización del término "descansar" afecta negativamente a las calificaciones de Didi Soho Hotel";

91. "La utilización del término "cocina" no afecta negativamente a las calificaciones de Didi Soho Hotel";

92. "La utilización del término "ubicación" no afecta negativamente y afecta positivamente a las calificaciones de Didi Soho Hotel";

93. "La utilización del término "habitación" afecta más positivamente de lo que afecta negativamente a las calificaciones de Didi Soho Hotel";

94. "La utilización de términos potenciadores del habla afectan más positivamente de lo que afecta negativamente a las calificaciones de Didi Soho Hotel";

95. "La utilización del término "cama" afecta más positivamente de lo que afecta negativamente a las calificaciones de Didi Soho Hotel";

96. "La utilización del término "atención" afecta más positivamente de lo que afecta negativamente a las calificaciones de Didi Soho Hotel";

97. "La utilización del término "ascensor" no afecta positivamente calificaciones de Didi Soho Hotel".

5.5. Capítulo 5: Clasificación por subjetividad

5.5.1. Introducción y definición de pautas de trabajo.

Según lo declarado por 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] Para esta tarea se utilizó la extensión MeaningCloud de Excel, y fue aplicada solamente al texto de las opiniones. Esta herramienta arroja dos categorías principales: "objetivo" (representando un lenguaje racional) y "subjetivo" (emocional o irracional), y en pocos casos "sin clasificación" (no fue posible su clasificación).

Recordando que la presente investigación contempla un contexto de ambigüedad, y por lo tanto, una situación de extrema incertidumbre, correspondió la realización de analogías (Bonatti et al, 2011, p. 76). Aun también habiendo tenido en cuenta las advertencias sobre la utilización de analogías, la cual "suele ser una de las fuentes más frecuentes de errores"; de hecho, podría mostrar qué no es certero, qué no deriva en un accionar o qué podría llegar a nublar la percepción de la realidad (Bonatti, 2005). Sin embargo, debido a que la presente investigación no intenta generar un modelo de decisión, ni establecer ninguna relación entre variables, sino que, uno de sus fines, es generar información a ser puesta a prueba en futuras investigaciones, se procedió con proponer un modelo análogo, explicado a continuación.

Para esta tarea se intentó relacionar los diferentes tipos de viajeros determinados por TripAdvisor) con los diferentes tipos de segmentación que señala el modelo de segmentación vincular (explicado más a fondo en el marco teórico), el cual trata de segmentar los diferentes tipos de vínculos entre el consumidor y el producto/servicio definiendo el vínculo en sí mismo entre estos dos (Wilensky, 2006; Caden, 1986). Particularmente, se hizo foco únicamente en señalar los tipos de viajeros que más podrían ajustarse al vínculo "racionalista"[38] (y por lo tanto al polo "discriminado"), o bien, al polo "simbiótico" (siendo sus posibles vínculos "comunitario" y/o "materno-filial") en un sentido amplio por no poder definir el vínculo más

[38] El vínculo "simbologista" fue descartado de este modelo análogo ya que al tratarse de un hotel 3 estrellas, se lo consideró muy poco probable el hecho de que alguien se hospede por cuestiones de prestigio o estatus social.

asociado a lo emocional. Por lo que, si un segmento demostró tener una clasificación de su vocabulario mayormente subjetiva, se lo asoció con el polo "simbiótico"; y si la clasificación de su vocabulario fue objetiva, se lo asoció con el vínculo "racional" del polo "discriminado".

En cuanto al valor "no especifica" de la variable "tipo de viajero", esta se la tuvo en cuenta solamente para controlar que la suma de todos los valores es igual a la muestra (100 unidades experimentales); por lo que de ninguna forma se lo tendrá en cuenta como un segmento del mercado turístico u hotelero. También, se descartaron otras variables que podrían haber sido puestas a prueba para comprobar si existe alguna relación entre sus diferentes valores y el resultado de la clasificación por subjetividad. Estas variables descartadas para este capítulo fueron: "puntaje del usuario", "fecha de alojamiento", "nivel de crítico", "utilización de dispositivo móvil", "clasificación por sentimiento", además de todas las otras variables descartadas desde el principio de la investigación (ver diseño muestral).

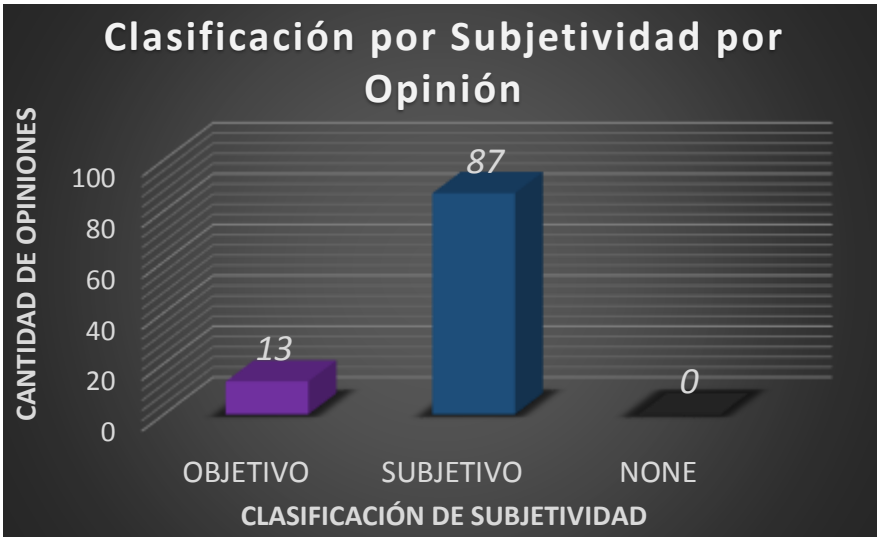
De esta forma se constituyó el modelo análogo correspondiente tendiente a establecer solamente una analogía. Debido que, como bien explica Wilensky (2006, p. 87), los seres humanos son seres complejos y no se sitúan en ninguno de los dos extremos; y además existen vínculos combinatorios en donde los cuatro vínculos puros explicados se entrelazan para formar otros tipos de vínculos (Wilensky, 2006, p. 92).

Para revisar si existe algún patrón de comportamiento entre los datos que pueda relacionarse con el modelo de segmentación vincular, se utilizó la función de tabla pivot que ofrece Excel, poniendo en el orden de filas a los tipos de viajeros y en el orden de columnas a las categorías según la clasificación por subjetividad de la herramienta. Además, estas tablas se crearon con dos tipos de valores que sirvieron para control mutuo entre las tablas, uno se trató de los valores numéricos, y el otro se trató del porcentaje. El primero se lo utilizó para saber si se está tratando con frecuencias muy pequeñas de las cuales pueda desembocar una variación muy grande en los porcentajes, de ser así, podría estar sospechándose de que se traten de valores atípicos. El segundo se lo utilizó principalmente para evidenciar algún patrón de comportamiento.

5.5.2. Clasificación a nivel opinión.

En una primera instancia, la tarea de clasificación por subjetividad se realizó para los conjuntos de opiniones de Didi Soho Hotel, siendo 100 opiniones en total. De esta se obtuvo un 87% de las opiniones clasificadas como subjetivas, un 13% como objetivas, y ninguna se clasificó como "no clasificable". Apréciese el siguiente gráfico de barras:

Figura 5.1. Clasificación por subjetividad por opinión



Fuente: producción propia en base a datos de Excel

Intuitivamente, parecería que se ha encontrado un comportamiento clave de la población, en donde la "gran mayoría", refiriéndose al 87% de las opiniones clasificadas como subjetivas, manifiesta un lenguaje subjetivo o emocional. Sin embargo, no debe olvidarse que Bing (2012, p. 45) expresa que "los textos subjetivos frecuentemente expresan puntos de vista y opiniones". Por lo que, se tuvo en cuenta, que de por sí, las opiniones son subjetivas, y se debe realizar una exploración más a fondo. [i]

Con el fin de poder encontrar algún indicio de información significativo, se procedió con seguir explotando los datos obtenidos y almacenados en forma de base de datos. Con estos datos se crearon diversas tablas pivot con Excel[39], y se obtuvieron los siguientes resultados cruzando la clasificación por subjetividad con otras variables, que son:

Figura 5.2. Clasificación por subjetividad por opinión según motivo de viaje

	Objetivo	Subjetivo	Total
Amigos	3	19	22
Familia	5	25	30
Negocios	1	15	16
No especifica	1	5	6
Pareja	3	19	22
Solo	0	4	4
Total	13	87	100

	Objetivo	Subjetivo	Total
Amigos	14%	86%	100%
Familia	17%	83%	100%
Negocios	6%	94%	100%
No especifica	17%	83%	100%
Pareja	14%	86%	100%
Solo	0%	100%	100%
Total	13%	87%	100%

Fuente: producción propia en base a datos de Excel

[39] Para los porcentajes se utilizó el formato condicional – escala de colores – verde/amarillo/rojo.

Según lo dispuesto en este par de tablas, se pudo ver que el segmento cuyo motivo del viaje fue pasar un momento en "familia" presentaron el mayor grado de objetividad en cuanto al número de opiniones (alrededor del 17%). Mientras que aquellos segmentos en donde los turistas viajaron entre "amigos" y en "pareja", cada segmento presentó un 14% de opiniones clasificadas como objetivas. Para el caso del segmento "negocios", solo 1 de 15 (aproximadamente el 6%) mostró un lenguaje predominante objetivo. Mientras que aquellos que viajaron mediante la categoría "solo", no presentaron ninguna clasificación objetiva, pero visto que solo hubo 4 comentarios en esta categoría, habría que considerar la posibilidad de que uno o más de estos 4 se traten de un valor atípico.

Teniendo estos valores, no fue posible construir ninguna proposición por parte del investigador, por lo que se procedió a utilizar la misma técnica de clasificación por subjetividad, pero con otros criterios de clasificación.

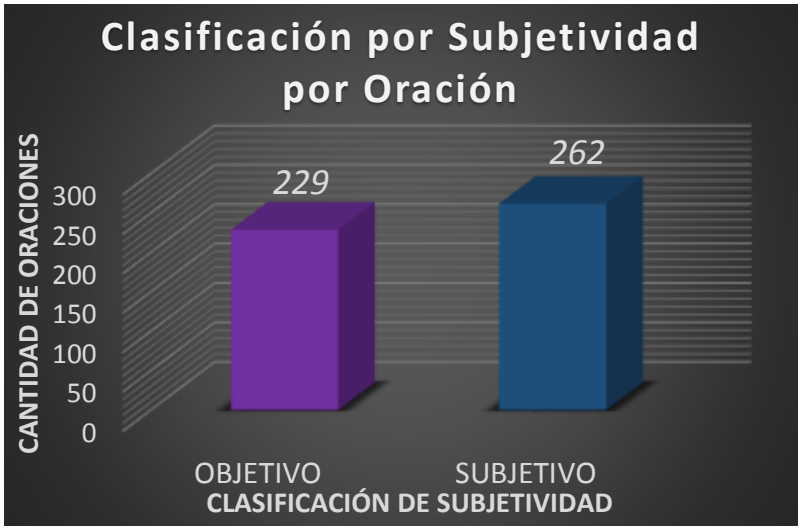
5.5.3. Clasificación a nivel oración sin discriminar opiniones.

Al ver que la clasificación por subjetividad a nivel opinión no deja en evidencia ningún comportamiento de la población que sea relevante ni identificable a simple vista, se procedieron con las recomendaciones de otros autores para que la clasificación por subjetividad arroje algún patrón identificable.

Según señalan Pang y Lee (2008, p. 29) [i], existen diversos proyectos en donde la clasificación se puede realizar tanto a nivel de la oración como a nivel de la sub-oración. Teniendo en cuenta esto, se procedió con descomponer las opiniones en oraciones. Para esta sub tarea, se utilizó R; las oraciones fueron determinadas según su finalización en "." (punto) y "!" (signo de exclamación). Acerca de las sub-oraciones, estas fueron omitidas. Principalmente, debido a que su descomposición desde la opinión requiere un mayor estudio de cómo se forman y funcionan en idioma español, además, una optimización más compleja en el código de programación ad hoc de la herramienta R.

De esta forma, se obtuvieron unas 491 oraciones de las 100 opiniones de Didi Soho Hotel. En donde, 262 oraciones (que conforman el 53,36% del total de opiniones del hotel) fueron clasificadas como subjetivas, las 229 oraciones restantes (46,64% del mismo total) fueron clasificadas como objetivas, y ninguna fue clasificada como "no clasificable". Apréciese el siguiente gráfico de barras:

Figura 5.3. Clasificación por subjetividad por oración



Fuente: producción propia en base a datos de Excel

Visto que, a comparación de la clasificación por subjetividad a nivel opinión, la clasificación a nivel oración mostró proporciones más equitativas, se procedió con producir la siguiente tabla pivot de acuerdo con las mismas especificaciones con las que se las realizaron a nivel opinión.

Figura 5.4. Clasificación por subjetividad por oración según motivo de viaje

	Subjetivo	Objetivo	Total
Amigos	57	65	122
Familia	74	68	142
Negocios	42	29	71
No especifica	14	9	23
Pareja	62	49	111
Solo	13	9	22
Total	262	229	491

	Subjetivo	Objetivo	Total
Amigos	47%	53%	100%
Familia	52%	48%	100%
Negocios	59%	41%	100%
No especifica	61%	39%	100%
Pareja	56%	44%	100%
Solo	59%	41%	100%
Total	53%	47%	100%

Fuente: producción propia en base a datos de Excel

Según lo expuesto por el conjunto de tablas, se pudo ver que el único segmento que superó la barrera de al menos un 50% de lenguaje objetivo en su conjunto, fue el segmento de viajeros entre "amigos" con un 53%, seguidos por el segmento "familia" con 48%, y luego el segmento "pareja" con 44%. Para los segmentos de "negocios" y "solo", estos solo presentaron un 41% de lenguaje objetivo, pero nuevamente se tuvo que tener en cuenta la posibilidad de que en el segmento "solo" exista algún valor atípico. Por lo que, se prefirió a partir de este punto, descartar el análisis del segmento "solo".

Quizás en este caso hubo alguna evidencia de algún patrón de comportamiento interesante para analizar, pero se prefirió proseguir con el siguiente criterio de clasificación, para ver si el patrón de comportamiento persiste.

5.5.4. Clasificación a nivel oración discriminadas por opinión.

Sabiendo que la tarea anterior tampoco ha dejado en evidencia ningún patrón de conducta de la población, se procedió con discriminar las opiniones sobre la clasificación a nivel oración, es decir, para cada oración se tuvo en cuenta a qué opinión pertenece.

Teniendo en cuenta que las opiniones pueden estar compuestas tanto de oraciones clasificadas como subjetivas como oraciones clasificadas como objetivas, se procedió con calcular un coeficiente de objetividad, en donde, para cada opinión, este coeficiente es el cociente entre la cantidad de oraciones clasificadas como objetivas, y la suma del total de las oraciones de la opinión (tanto objetivas como subjetivas). Por lo tanto, será expresada mediante la siguiente fórmula:

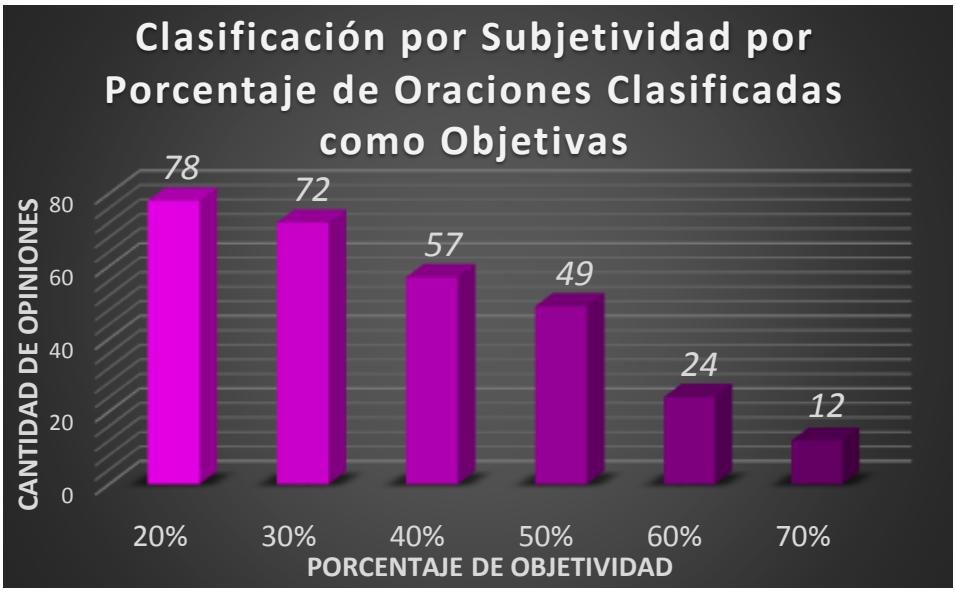
$$Coeficiente\ de\ Objetividad_{[1;100]} = \frac{Q\ oraciones\ objetivas_{[1;100]}}{Q\ total\ oraciones_{[1;100]}}$$

Una vez obtenido este coeficiente para cada una de las oraciones, se procedió con clasificar cada opinión en 6 (seis) variables dicotómicas, en donde los valores pueden ser "0" (cero) para falso y "1" (uno) para verdadero[40]. 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%[41]. 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%. De esta forma, se consiguieron crear tanto el siguiente gráfico a continuación, como la tabla pivot entre los segmentos y las categorías para el coeficiente de objetividad, con las mismas especificaciones que las tareas de clasificación por subjetividad anteriores.

[40] Para calcular el coeficiente de objetividad y clasificar a las opiniones en las distintas categorías se utilizó el lenguaje R.

[41] Las categorías de 0%, 10%, 80%, 90% y 100%, fueron descartadas ya que en etapas previas de experimentación y confección del código en R, 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 5.5. Clasificación por subjetividad por porcentaje de oraciones clasificadas como objetivas



Fuente: producción propia en base a datos de Excel

Figura 5.6. Clasificación por subjetividad por porcentaje de oraciones clasificadas como objetivas según el motivo de viaje

	70%	60%	50%	40%	30%	20%	Total
Amigos	4	9	15	17	18	20	22
Familia	3	7	18	19	26	26	30
Negocios	1	3	6	9	11	12	16
No especifica	1	1	1	1	2	3	6
Pareja	2	3	8	10	13	14	22
Solo	1	1	1	1	2	3	4
Total	12	24	49	57	72	78	100

	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: producción propia en base a datos de Excel

Para este caso, vemos de vuelta al segmento de “amigos” ocupar la cabeza, en donde se pudo ver que un 77% del lenguaje utilizó al menos un 40% de lenguaje objetivo para describir su experiencia en el hotel, 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, pero siguiendo a la cabeza, 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 10% 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 se encuentran relativamente cerca, más que los segmentos de “amigos” y “familia”, en donde a partir de la categoría de “50%”, se puede ver que sus

coeficientes se aproximan a la mitad de los segmentos que están a la cabeza. En cuanto a los segmentos "no especifica" y "solo" fueron descartados según lo dispuesto en puntos anteriores.

Teniendo la información generada a lo largo de esta parte, fue posible exponer las siguientes proposiciones:

98. *"El segmento de viajeros entre amigos que busca hoteles 3 estrellas en el barrio de Palermo, es el que más presenta un vínculo racionalista con el producto hotelero, y por lo tanto están en un polo discriminado";*
99. *"El segmento de familia es ligeramente menos racionalista que el segmento de viajeros entre amigos que busca hoteles 3 estrellas en el barrio de Palermo, pero sigue estando más en el polo discriminado que en el polo simbiótico";*
100. *"El segmento de familia se sitúa ligeramente más en el polo simbiótico que el segmento de viajeros entre amigos";*
101. *"Los segmentos de viajeros de negocios y parejas presentan un menor vínculo racional con el producto hotelero que los segmentos de amigos y familia, y por lo tanto es más ajustable al polo simbiótico";*
102. *"El posicionamiento en el mercado hotelero argentino para los segmentos de viajeros entre amigos y familia se establece por diferenciación por precio";*
103. *"El posicionamiento en el mercado hotelero argentino para los segmentos de viajeros de negocios y parejas se establece por diferenciación por calidad";*
104. *o bien, "El posicionamiento en el mercado hotelero argentino para los segmentos de viajeros de negocios y parejas se establece por diferenciación por marca"* [42].

[42] Wilensky (2006) menciona repetidas veces la relación entre las marcas con lo simbólico y la identidad, el cual se entendería que está relacionado con el vínculo "simbologista", el cual fue descartado en este trabajo. Sin embargo, se asumió que la lealtad a la marca posee una menor carga racional que la que posee el vínculo "racionalista", por lo que la marca podría ser, un motor de elección del producto hotelero para los demás segmentos que están menos vinculados con lo racional y más vinculados con lo emocional.

5.6. Capítulo 6: Generación de tópicos

5.6.1. Introducción y definición de las pautas de trabajo.

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. Por ejemplo, en el análisis del sentimiento, los tópicos son aspectos en el contexto del análisis del sentimiento, por lo que diversos modelos y técnicas pueden utilizarse para extraer aspectos analizando el contenido del documento. [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";
  - "habitaciones".
- Además, se agregaron los siguientes temas para poder contar con más variables:
  - "equipamiento": se trata de todos aquellos objetos que posee la habitación, más allá de las características inherentes de esta como lo puede ser el tamaño, la comodidad y el estado de mantenimiento de 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": se trata principalmente de 3 cuestiones 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[43], 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[44]. 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.

5.6.2. Clasificación de opiniones según sus tópicos.

Figura 6.1. Cantidad de opiniones por tópicos

Fuente: producción propia en base a datos de Excel.0

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)

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

[44] Ver conjuntos de términos utilizados para la identificación de aspectos en la sección de Anexos.

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). Esto dio lugar a las siguientes proposiciones por parte del investigador:

105. "Tanto las 'habitaciones' como la 'ubicación' son los aspectos que más suelen importarle al huésped a la hora de evaluar el producto hotelero";

106. "Tanto la 'limpieza' como la calidad de 'atención' son aspectos que quedan en segundo plano a la hora de evaluar el producto hotelero";

107. "Otros aspectos como la 'relación precio-calidad', la 'calidad de descanso', el 'equipamiento' de las habitaciones y las facilidades en cuanto a la 'tecnología' son de menor importancia para el huésped a la hora de evaluar un producto hotelero".

5.6.3. Clasificación de opiniones sin discriminar sus tópicos.

Por otro lado, se procedió también con 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, a fin de conocer más acerca de la población que compone el hotel y/o la totalidad de los huéspedes en la comunidad TripAdvisor.com.

De esta forma, entonces, se prosiguió con realizar el conteo con la totalidad de los tópicos generados hasta el momento por el investigador, y se obtuvo el siguiente gráfico creado con Excel y utilizando los resultados que arroja R, seguidos de sus respectivos atributos de estadística descriptiva (también calculados mediante R):

Figura 6.2. Cantidad tópicos no discriminados por opiniones

Fuente: producción propia en base a datos de Excel.

Figura 6.3. Medidas de estadística descriptiva de la cantidad de tópicos no discriminados por opinión

	Didi Soho Hotel	Observaciones
$\bar{x}$	4.13	
σ	1.236278	
Me(x)	4	
Mo(x)	4	29 repeticiones
Coeficiente de asimetría	0.20603	La distribución es asimétrica hacia la derecha
Coeficiente de curtosis	-0.12521	La distribución es platicúrtica, pero casi mesocúrtica

Fuente: producción propia en base a datos de R.

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

108. "Los turistas que se hospedan en hoteles 3 estrellas del barrio de Palermo suelen percibir entre 3 a 5 aspectos del producto turístico";

109. "La cantidad de aspectos a la que un turista o huésped suele hacer referencia es una variable que posee una distribución normal".

5.7. Capítulo 7: Reglas de asociación

5.7.1. Introducción a la técnica.

Según lo establecido por Witter et al. (2011) las reglas de asociación sirven para predecir el valor de una variable, dado el comportamiento de una o más variables. [i]

Para esta tarea en particular se precisó que algunas de las técnicas anteriores fueran completadas (clasificación por subjetividad, clasificación por análisis del sentimiento, y generación de tópicos), ya que fue el conjunto de información generado por esas técnicas el que fue sometido al experimento de esta técnica en particular.

Además, fue necesario realizar algunas modificaciones a las variables tratadas a lo largo del trabajo, lo cual se verá a continuación

5.7.2. Creación y definición de variables.

Debido a que solamente se contó con una muestra de 100 unidades, se previó que las reglas de asociación no iban a ser lo suficientemente potentes de establecer reglas y detectar algún tipo de comportamiento significativo, ya que las variables poseen gran cantidad de categorías y sería muy poco probable que dos categorías de dos variables diferentes pudieran estar relacionadas. Por este motivo, se procedió con reducir las categorías de cada variable para que las reglas, si bien terminaron siendo más generales y menos refinadas, se pudo contar con reglas para observar la asociación entre las categorías de las variables.

Las variables modificadas para tales experimentos y cómo se han agrupado las categorías fueron hechas de la siguiente manera:

Figura 7.1. Modificación de variables mediante reducción del número de categorías

Variable Original	Categoría Original	Categorías Agrupadas	Criterios
Nivel de crítico	0	Bajo	El salto más grande entre las categorías, según TripAdvisor, está entre las categorías 3 y 4, en donde para ser crítico nivel 4 se le exigen 1.500 puntos, un 150% más que al crítico nivel 3, la cual solamente exige 1.000 puntos, un 100% más que al crítico nivel 2 (500 puntos mínimo).
	1		
	2		
	3		
	4	Alto	
	5		
	6		
Calificación del usuario	1	Malo	Debido a que Didi Soho Hotel es un hotel 3 estrellas y visto a que sus medidas de tendencia central están entre los puntos 3 y 4, se estableció esta línea divisoria entre los puntajes 3 y 4.
	2		
	3	Bueno	
	4		
	5		

Fecha de alojamiento (meses)	Diciembre	12	Dic – Feb	Estas son las divisiones que ofrece TripAdvisor en el filtro llamado "época del año", el cual toma períodos de cada 3 meses corridos del año.
	Enero	1		
	Febrero	2		
	Marzo	3	Mar – May	
	Abril	4		
	Mayo	5		
	Junio	6	Jun – Ago	
	Julio	7		
	Agosto	8		
	Septiembre	9	Sep – Nov	
	Octubre	10		
	Noviembre	11		
Fecha de alojamiento (años)	2017		2016 – 2017	La división 2016-2017 y 2014-2015 es la que más equitativa deja la cantidad de opiniones para cada categoría.
	2016			
	2015		2014 – 2015	
	2014			
Clasificación por sentimiento	N+	1	Negativo	Proveniente del capítulo 4 (análisis del sentimiento) al igual que el criterio para la calificación del usuario, se optó por catalogar a la clasificación por sentimiento con valor 3 en la categoría "positivo".
	N	2		
	NEU	3	Positivo	
	P	4		
	P+	5		
Cantidad de tópicos	0		2 o menos; o, 6 o más	De acuerdo con el capítulo 6: Generación de tópicos, la mayoría de las opiniones hacen referencia a entre 3 y 5 tópicos.
	1			
	2			
	3		3 a 5	
	4			
	5			
	6		2 o menos; o, 6 o más	
	7			
	8			

Fuente: producción propia con variables tratadas a lo largo del trabajo.

Además, también se utilizaron las variables dicotómicas generadas en el capítulo 5 (Clasificación por subjetividad), estas son: "clasificación por subjetividad" a nivel opinión, y la "clasificación por subjetividad" por porcentajes a nivel oración discriminando la opinión. De esta forma, se suman 7 (siete) variables dicotómicas para este capítulo de reglas de asociación provenientes del capítulo 5.

Del capítulo 6 (Generación de tópicos), se incluyeron los 8 (ocho) tópicos generados en forma de variable dicotómica, ya que debe indicar si el usuario se refirió a este tópico o no; y además la novena variable (ver tabla de creación de variables) que indica si la opinión se refirió a entre 3 y 5 aspectos en su opinión, teniendo en cuenta la supuesta normalidad en la distribución de esta variable formulada en el capítulo 6: Generación de tópicos.

Entre todas las variables utilizadas se logró definir 23 variables, de las cuales:

- 20 variables dicotómicas:

    - Variables que ofrece TripAdvisor (4): “nivel de crítico”, “calificación”, “año”, “utilización de dispositivo mobile”;
    - Del capítulo 4: 1 variable dicotómica;
    - Del capítulo 5: 7 variables dicotómicas;
    - Del capítulo 6: 9 variables dicotómicas;
- 1 variable de 4 categorías nominales:
    - “meses” (agrupadas en épocas del año), proveniente de TripAdvisor;
- y 1 variable de 5 categorías nominales:
    - “motivo del viaje” (original), proveniente de TripAdvisor.

5.7.3. Aplicación de la técnica sin definir variables.

5.7.3.1. Primer abordaje.

Una vez que se definieron las variables, se utilizó R[45] para tomar tanto los datos recolectados en la etapa de muestreo como la información generada en los capítulos anteriores, algunos con ciertas modificaciones especificadas, para conformar un único archivo .csv que contenga los datos de ambas fuentes.

En un principio, si no se especificara ningún tipo de criterio para descubrir las reglas de asociación que permitió el conjunto de variables, la herramienta[46] arrojó en total unas 3.027.317 reglas posibles. Por lo que fueron especificados los siguientes criterios para reducir esta gran cantidad de reglas:

- Solo se utilizaron 2 variables por regla (1 de contexto y 1 de resultado);
- Se eliminaron las reglas redundantes[47];
- Solo se tuvieron en cuenta aquellas reglas con un nivel de confianza del 80% o más.

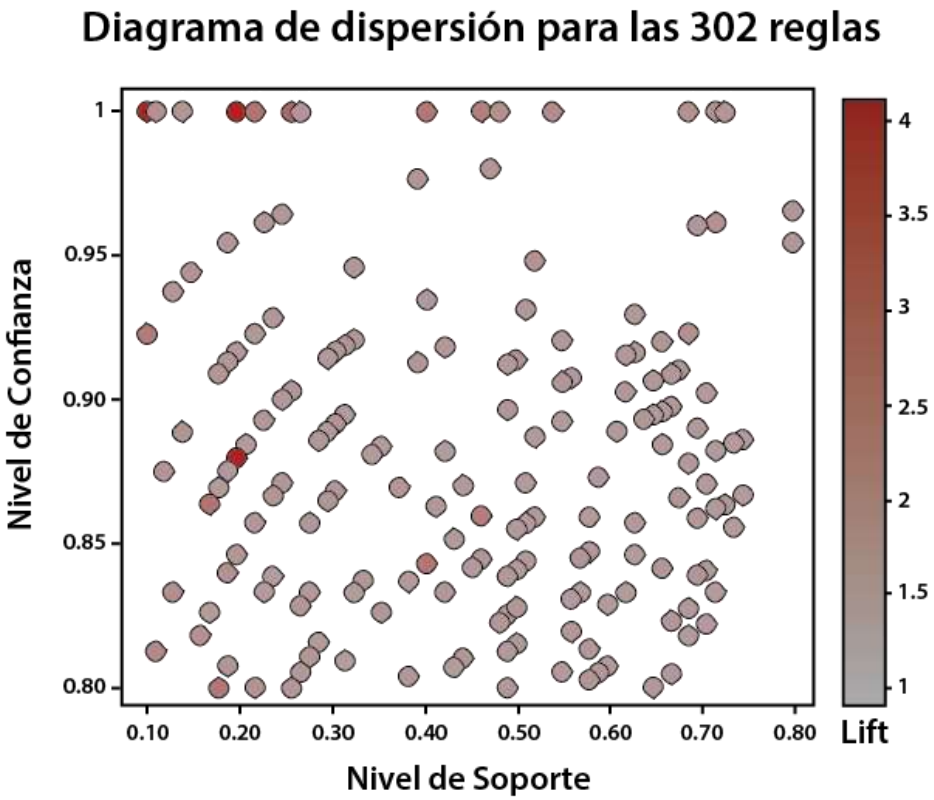
[45] Principales paquetes utilizados: “arules”, “arulesViz”, “kernlab”.

[46] Se utilizó el algoritmo “a priori”, el cual no puede crear reglas de asociación con más de 10 variables involucradas, utiliza un nivel de confianza del 80% como mínimo y un nivel de soporte del 10% como mínimo, en cuanto al “lift” no posee ni límite inferior ni límite superior.

[47] Las reglas de asociación redundantes son aquellas reglas cuyas variables utilizadas son exactamente las mismas de otra regla, con la única diferencia de que cambia el orden de utilización de las variables.

Una vez hecho esto, se logró reducir el conjunto de reglas generadas a 302 reglas de asociación bajo estos criterios; por lo que se podría suponer que, mediante esta técnica, al menos unas 302 hipótesis pudieron ser generadas[48]. Estas reglas fueron representadas mediante el siguiente diagrama de dispersión[49]:

Figura 7.2. Diagrama de dispersión para las reglas generadas



Fuente: producción propia en base a datos de R.

En este diagrama se puede apreciar que las reglas con mayor nivel de "lift" están en donde el nivel de soporte se aproxima al eje de ordenadas. Por lo que se podría suponer que el nivel de "lift" es propenso a aumentar a medida en que el nivel de soporte se reduce. Dicho esto, se pudo proponer[50] lo siguiente con respecto a las posibles reglas de asociación (con dos variables) generadas para Didi Soho Hotel:

[48] Sin contar que cada regla posee 3 formas de medición: soporte, nivel de confianza y "lift".

[49] La función de dibujo del diagrama de reglas de asociación sitúa por defecto al soporte en el eje de abscisas, al nivel de confianza en el eje de ordenadas, y el nivel de "lift" es representado por la intensidad del color de cada punto (que representa una regla de asociación).

[50] Algunas palabras, principalmente las de connotación positiva y negativa fueron escritas con letras mayúsculas para facilitar la lectura y la comprensión por parte del lector.

110. "Existe relación INVERSA o relación NEGATIVA entre el nivel de 'lift' y el soporte";

111. "NO existe relación entre el nivel de confianza y el nivel de 'lift'";

112. "Existe relación INVERSA o relación NEGATIVA entre la cantidad de reglas de asociación generadas y el nivel de confianza";

113. "A MAYOR nivel de soporte, MENOR nivel de confianza".

5.7.3.2. Segundo abordaje.

Por otro lado, habiendo revisado las reglas generadas[51], aquellas que posee los niveles de confianza más altos. Las primeras reglas obtenidas fueron la siguiente:

Figura 7.3. Reglas generadas según nivel de confianza

Antecedente		Consecuente	Soporte	Confianza	Lift
{sentimiento=negativo}	=>	{calificacion=malo}	21.78%	100%	4.04
{objetividad70=1}	=>	{objetividad60=1}	11.88%	100%	4.21
{objetividad70=1}	=>	{objetividad50=1}	11.88%	100%	2.06
{objetividad70=1}	=>	{objetividad40=1}	11.88%	100%	1.77
{objetividad70=1}	=>	{objetividad30=1}	11.88%	100%	1.40
{objetividad70=1}	=>	{objetividad20=1}	11.88%	100%	1.29
{objetividad70=1}	=>	{aspecto_habitacion=1}	11.88%	100%	1.12
{subjetividad=0}	=>	{objetividad20=1}	12.87%	100%	1.29
{subjetividad=0}	=>	{aspecto_habitacion=1}	12.87%	100%	1.12
{motivo=negocios}	=>	{aspecto_tecnologia=0}	15.84%	100%	1.19

Fuente: producción propia en base a datos de R.

Tomando la primera regla como ejemplo, a simple vista se pudieron constituir las siguientes proposiciones con los 3 tipos de mediciones:

114. "Cuando la clasificación del sentimiento es 'NEGATIVO' se puede afirmar con un nivel de confianza del 100% de que la calificación fue 'MALO'";

115. "El 22% de la muestra afirma que cuando la clasificación del sentimiento es 'NEGATIVO', la calificación fue 'MALO'";

[51] Solo se exponen las primeras 10 reglas en la lista. Para ver las primeras 100 de las 302 reglas de asociación generadas dirigirse a la sección de Anexos.

116. *"Existe una confianza en la mejora de 4 de que cuando la clasificación del sentimiento es 'NEGATIVO', la calificación fue 'MALO'".*

Pero además, utilizando las otras categorías de las variables involucradas, también se puede deducir lo siguiente:

- Cambiando la categoría de la variable antecedente:

117. *"Cuando la clasificación del sentimiento dio 'POSITIVO', se pudo exponer con un nivel de confianza DISTINTO al 100% de que la calificación dio 'MALO';*

- O bien, cambiando la categoría de la variable consecuente:

118. *"Cuando la clasificación del sentimiento dio 'NEGATIVO', se pudo exponer con un nivel de confianza DISTINTO al 100% de que la calificación fue 'BUENO'";*

- O bien, cambiando las categorías de ambas variables:

119. *"Cuando la clasificación del sentimiento es 'POSITIVO' se puede afirmar con un nivel de confianza del 100% de que la calificación fue 'BUENO'";*

120. *O también, "Cuando la clasificación del sentimiento es 'POSITIVO' se puede afirmar con un nivel de confianza DISTINTO al 100% de que la calificación fue 'BUENO'".*

Por lo tanto, de una sola regla de asociación, se pueden deducir 7 proposiciones que podrían servir de información y/o como hipótesis. Sin embargo, esta forma de leer e interpretar las primeras reglas que la técnica arroja, no resulta muy práctico a la hora de ordenar la información y que ayude a entender cómo se comportan los datos. Por lo que versiones más focalizadas fueron tratadas en los siguientes puntos de esta parte.

5.7.4. Aplicación definiendo variables.

Visto que las reglas de asociación generadas poseen una gran cantidad de reglas que relacionan los distintos niveles de objetividad (20%, 30%, etc.) y la clasificación por subjetividad situándolos en los niveles más altos de confianza, esto no mostró ninguna clase de información

relevante[52]. Por este motivo, se optó, en los puntos siguientes, determinar las diferentes variables con el fin de obtener una mirada más profunda en los datos y poder descubrir ciertos patrones de comportamiento en estos datos.

Las reglas de asociación generadas fueron ordenadas de forma descendente por su nivel de confianza. Las representaciones se realizaron mediante las matrices de correlación que ofrece R y fueron modificadas con Adobe Illustrator para mejorar las posibilidades de interpretación.

5.7.4.1. Reglas de asociación determinando las variables antecedentes.

En este punto se intentó establecer reglas de asociación determinando cuáles son las variables antecedentes junto con sus valores, es decir, teniendo la información de ciertas variables, cuáles variables junto con sus valores fueron los que más ocurrieron.

Se estableció como variable antecedente al "motivo de viaje", en donde las posibles categorías a probar fueron: "amigos", "negocios", "familia" y "pareja", la categoría "solo" se la consideró muy pequeña como para ser tenida en cuenta, y la categoría "no especifica" se aclaró anteriormente que solo se la tuvo en cuenta para controlar que la muestra sea igual a la establecida en la etapa de muestreo. De esta forma, se obtuvieron las siguientes 22 reglas de asociación[53]:

Figura 7.4. Lista de reglas generadas con el motivo de viaje como variable antecedente

Antecedente		Consecuente	Soporte	Confianza	Lift
{motivo=negocios}	=>	{aspecto_tecnologia=0}	16%	100.00%	1.19
{motivo=pareja}	=>	{aspecto_habitacion=1}	22%	100.00%	1.11
{motivo=amigos}	=>	{aspecto_habitacion=1}	21%	95.45%	1.06
{motivo=negocios}	=>	{aspecto_ubicacion=1}	15%	93.75%	1.14
{motivo=negocios}	=>	{subjetividad=1}	15%	93.75%	1.08
{motivo=negocios}	=>	{objetividad70=0}	15%	93.75%	1.07
{motivo=amigos}	=>	{objetividad20=1}	20%	90.91%	1.17

[52] Ya que no debería considerarse como información que: aquellas opiniones con un 70% de su vocablo clasificado como objetivo, posea un 100% de nivel de confianza que también utilice un 60% de su vocablo como objetivo; lo mismo funciona entre el 60% y el 50%, el 50% y el 40%, etc. Es fácilmente deducible por lógica, y no se necesita la herramienta o los datos para ello. Si estos resultados no dieran un nivel de confianza del 100%, algo estaría mal con las herramientas.

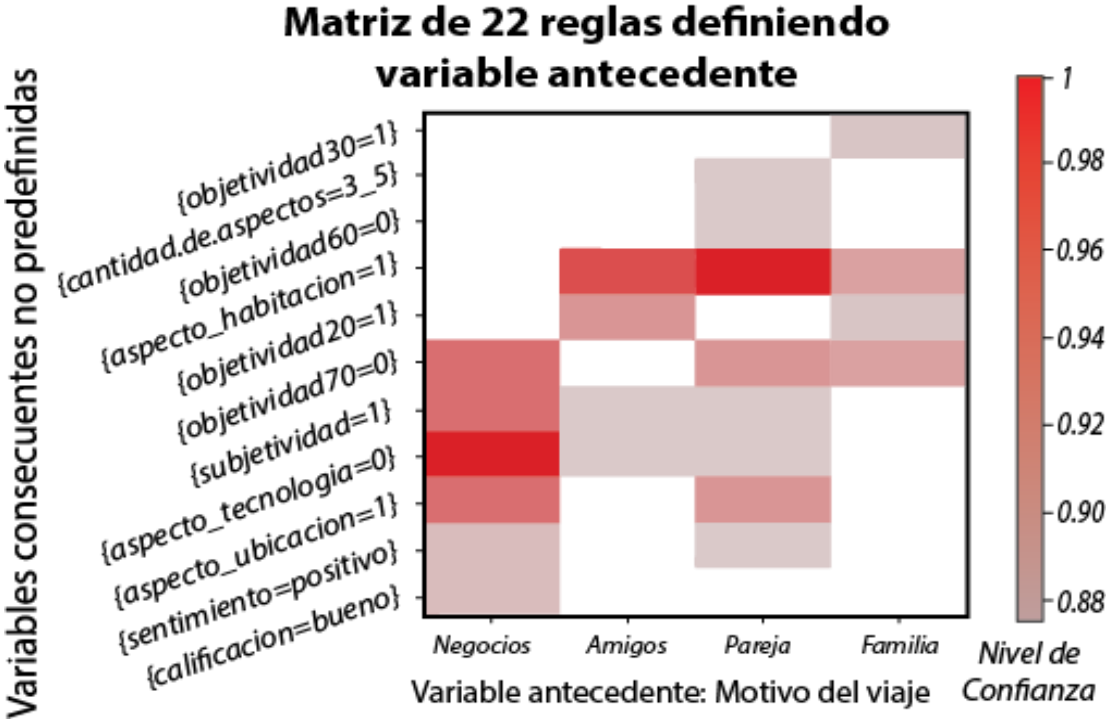
[53] Se especificó un nivel de confianza como mínimo del 85%, y el resto de los criterios fueron los mismos que el punto anterior.

{motivo=pareja}	=>	{aspecto_ubicacion=1}	20%	90.91%	1.11
{motivo=pareja}	=>	{objetividad70=0}	20%	90.91%	1.03
{motivo=familia}	=>	{objetividad70=0}	27%	90.00%	1.02
{motivo=familia}	=>	{aspecto_habitacion=1}	27%	90.00%	1.00
{motivo=negocios}	=>	{calificacion=bueno}	14%	87.50%	1.17
{motivo=negocios}	=>	{sentimiento=positivo}	14%	87.50%	1.12
{motivo=familia}	=>	{objetividad30=1}	26%	86.67%	1.20
{motivo=familia}	=>	{objetividad20=1}	26%	86.67%	1.11
{motivo=amigos}	=>	{aspecto_tecnologia=0}	19%	86.36%	1.03
{motivo=amigos}	=>	{subjetividad=1}	19%	86.36%	0.99
{motivo=pareja}	=>	{objetividad60=0}	19%	86.36%	1.14
{motivo=pareja}	=>	{cantidad.de.aspectos=3_5}	19%	86.36%	1.12
{motivo=pareja}	=>	{sentimiento=positivo}	19%	86.36%	1.11
{motivo=pareja}	=>	{aspecto_tecnologia=0}	19%	86.36%	1.03
{motivo=pareja}	=>	{subjetividad=1}	19%	86.36%	0.99

Fuente: producción propia en base a datos de R.

Las cuales pueden ser representadas mediante el siguiente gráfico:

Figura 7.5. Matriz de reglas generadas con el motivo de viaje como variable antecedente



Fuente: producción propia en base a datos de R.

Aquí se vio cómo los distintos segmentos son más propensos que otros para:

- Utilizar o no utilizar una mayor parte de su vocabulario como objetivo;

- Mostrar o no una clasificación más ajustada a la subjetividad;
- Calificar mejor o peor que otros;
- Mostrar un sentimiento más positivo o más negativo que otros;
- Situarse o no en la supuesta normalidad de incluir entre 3 y 5 aspectos en la opinión;
- Referirse o no a ciertos aspectos del producto hotelero.

Aún así, habiendo generado información relevante, se prosiguió con realizar otra amplificación en cuanto a los datos, y esta vez determinando ambas variables antecedentes (siendo esta el "motivo del viaje") y consecuentes.

5.7.4.2. Reglas de asociación determinando las variables antecedentes y consecuentes.

Con el fin de conocer más acerca del segmento, se optó por determinar, además de la variable antecedente como "motivo del viaje" (solo las 4 categorías que no se descartaron hasta el momento), y como variable consecuente, se consideró interesante el poder visualizar los diferentes aspectos o tópicos generados (los 8 del capítulo anterior y la 9na variable que señala la cantidad de tópicos). De esta forma se obtuvieron las siguientes 24 reglas de asociación[54]:

Figura 7.6. Lista de reglas generadas con el motivo de viaje como variable antecedente y los aspectos generados como variables consecuentes

Antecedente		Consecuente	Soporte	Confianza	Lift
{motivo=negocios}	=>	{aspecto_tecnologia=0}	16%	100.00%	1.19
{motivo=pareja}	=>	{aspecto_habitacion=1}	22%	100.00%	1.11
{motivo=amigos}	=>	{aspecto_habitacion=1}	21%	95.45%	1.06
{motivo=negocios}	=>	{aspecto_ubicacion=1}	15%	93.75%	1.14
{motivo=pareja}	=>	{aspecto_ubicacion=1}	20%	90.91%	1.11
{motivo=familia}	=>	{aspecto_habitacion=1}	27%	90.00%	1.00
{motivo=pareja}	=>	{cantidad.de.aspectos=3_5}	19%	86.36%	1.12
{motivo=pareja}	=>	{aspecto_tecnologia=0}	19%	86.36%	1.03
{motivo=amigos}	=>	{aspecto_tecnologia=0}	19%	86.36%	1.03
{motivo=amigos}	=>	{cantidad.de.aspectos=3_5}	18%	81.82%	1.06
{motivo=negocios}	=>	{aspecto_equipamiento=0}	13%	81.25%	1.14
{motivo=familia}	=>	{aspecto_descanso=0}	24%	80.00%	1.23
{motivo=amigos}	=>	{aspecto_equipamiento=0}	17%	77.27%	1.09

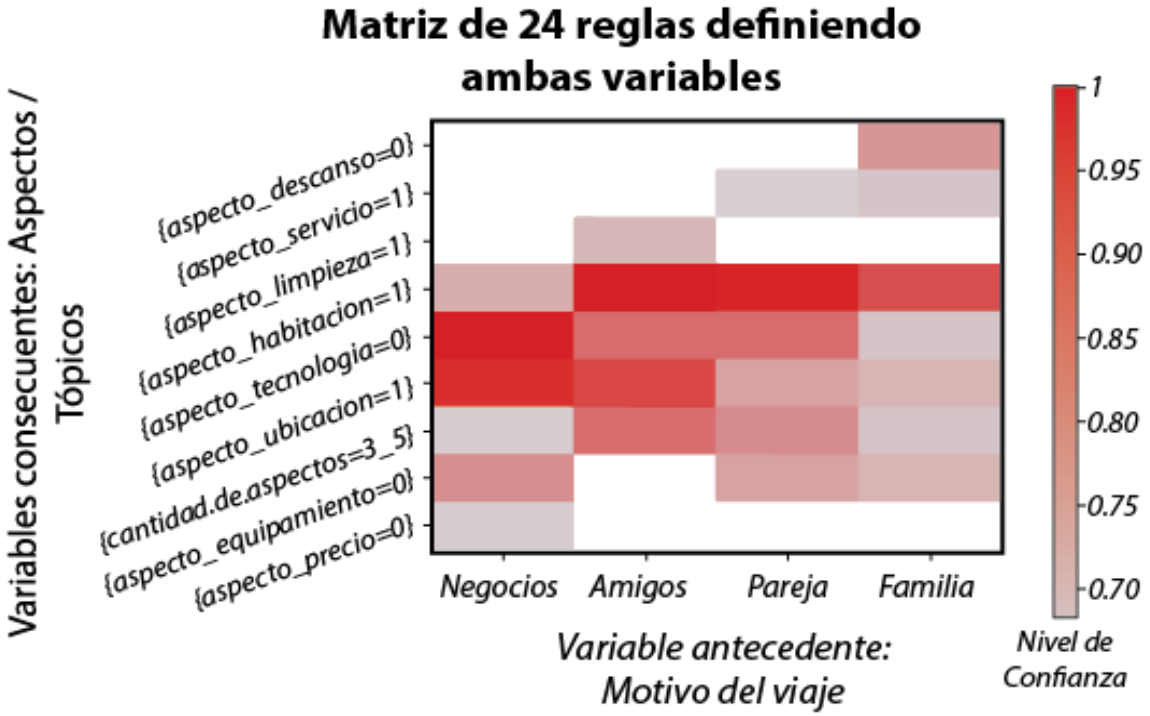
[54] Se especificó un nivel de confianza como mínimo del 65% y el resto de los criterios fueron los mismos que en el punto anterior.

{motivo=amigos}	=>	{aspecto_ubicacion=1}	17%	77.27%	0.94
{motivo=negocios}	=>	{aspecto_habitacion=1}	12%	75.00%	0.83
{motivo=familia}	=>	{aspecto_equipamiento=0}	22%	73.33%	1.03
{motivo=familia}	=>	{aspecto_ubicacion=1}	22%	73.33%	0.89
{motivo=pareja}	=>	{aspecto_limpieza=1}	16%	72.73%	1.15
{motivo=familia}	=>	{aspecto_servicio=1}	21%	70.00%	1.13
{motivo=familia}	=>	{cantidad.de.aspectos=3_5}	21%	70.00%	0.91
{motivo=familia}	=>	{aspecto_tecnologia=0}	21%	70.00%	0.83
{motivo=negocios}	=>	{aspecto_precio=0}	11%	68.75%	1.07
{motivo=negocios}	=>	{cantidad.de.aspectos=3_5}	11%	68.75%	0.89
{motivo=amigos}	=>	{aspecto_servicio=1}	15%	68.18%	1.10

Fuente: producción propia en base a datos de R.

Así, entonces se obtuvo el siguiente gráfico:

Figura 7.7. Matriz de reglas generadas con el motivo de viaje como variable antecedente y los aspectos generados como variables consecuentes



Fuente: producción propia en base a datos de R.

De esta versión, se puede obtener una aproximación más nítida acerca de cuáles son los tópicos hacia los cuales los diferentes segmentos suelen hacer una mayor referencia; que para este caso se pudo proponer, a grandes rasgos, lo siguiente:

121. *"Al segmento 'negocios' el aspecto que MÁS le importa es la 'ubicación'";*

122. *"A los segmentos 'amigos', 'pareja', y 'familia' el aspecto que MENOS le importa es la 'habitación'";*

123. *"El segmento 'amigos' es el que MÁS se ajusta a la supuesta normalidad de hacer referencia a entre 3 y 5 aspectos";*

124. *"El segmento 'familia' es el que MENOS se ajusta a la supuesta normalidad de hacer referencia a entre 3 y 5 aspectos";*

125. *"Al segmento 'amigos' es al que MÁS le interesa el aspecto 'limpieza'";*

126. *"Al segmento 'familia' es al que MÁS le interesa el aspecto 'servicio/atención'";*

127. *"Al segmento 'negocios' es al que MENOS le interesa el aspecto 'habitación'";*

128. *"Al segmento 'familia' es al que MENOS le interesa el aspecto 'descanso'";*

129. *"Al segmento 'negocios' es al que MENOS le interesa el aspecto 'precio'".*

Viendo la posibilidad de poder manipular las categorías de los dos tipos de variables, se procedió con determinar las variables consecuentes en el siguiente punto.

5.7.4.3. Reglas de asociación determinando las variables consecuentes.

Para esta ocasión, se optó por generar reglas de asociación cuya variable consecuente sea la "calificación del usuario", a fin de obtener pistas sobre las posibles causas, características o condiciones por las cuales la calificación del usuario es "malo" o "bueno". De esta forma, se obtuvieron las siguientes reglas[55]:

Figura 7.8. Lista de reglas generadas con la calificación como variable consecuente

Antecedente		Consecuente	Soporte	Confianza	Lift
{sentimiento=negativo}	=>	{calificacion=malo}	22%	100.00%	4.00
{sentimiento=positivo}	=>	{calificacion=bueno}	75%	96.15%	1.28
{anno=2014-2015}	=>	{calificacion=bueno}	53%	91.38%	1.22
{aspecto_limpieza=0}	=>	{calificacion=bueno}	33%	89.19%	1.19
{mes=dic-feb}	=>	{calificacion=bueno}	16%	88.89%	1.19
{motivo=negocios}	=>	{calificacion=bueno}	14%	87.50%	1.17
{aspecto_ubicacion=1}	=>	{calificacion=bueno}	69%	84.15%	1.12
{nivel.de.critico=alto}	=>	{calificacion=bueno}	38%	82.61%	1.10

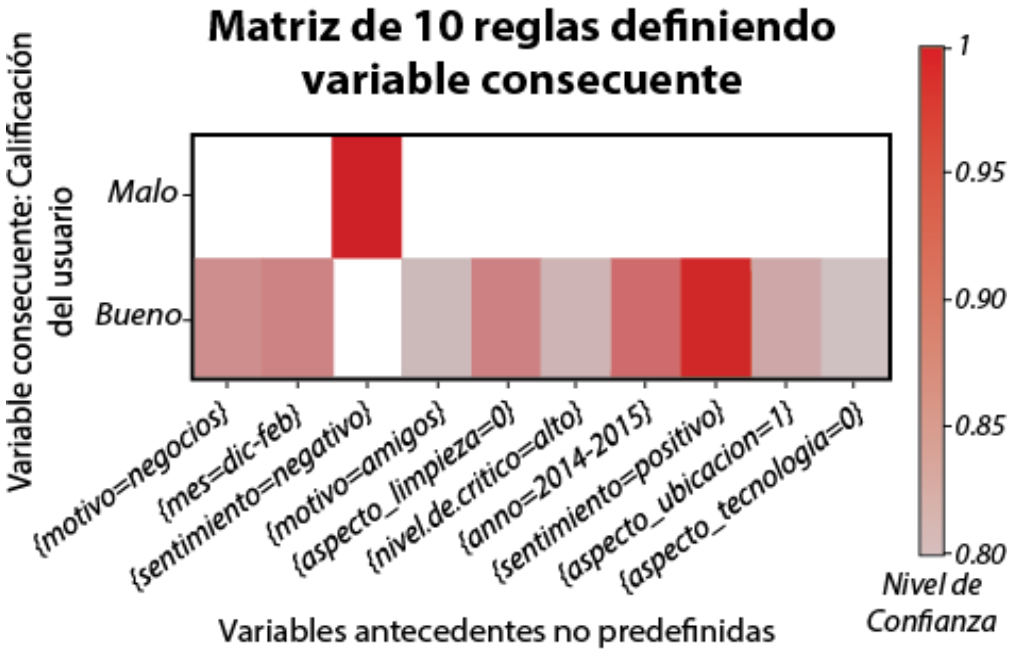
[55] Para ello, se optó por utilizar el mismo algoritmo que el paso anterior, el "a priori" para generar las reglas, al igual que los mismos criterios anteriores.

{motivo=amigos}	=>	{calificacion=bueno}	18%	81.82%	1.09
{aspecto_tecnologia=0}	=>	{calificacion=bueno}	68%	80.95%	1.08

Fuente: producción propia en base a datos de R.

A su vez, la siguiente imagen puede facilitar la visualización e interpretación de los datos:

Figura 7.9. Matriz de reglas generadas con la calificación como variable consecuente



Fuente: producción propia en base a datos de R.

De esta se pudieron deducir las siguientes proposiciones:

130. *"Las calificaciones que dieron 'BUENO' están fuertemente relacionadas con una clasificación del sentimiento POSITIVA";*

131. *"Las calificaciones que dieron 'BUENO' correspondieron más a una cierta época del año, y también, a conjunto de años";*

132. *De esta, también se puede deducir que "Las calificaciones que dieron 'MALO' también correspondieron más a una cierta época del año, y también, a conjunto de años";*

133. *"Ciertos segmentos tuvieron la tendencia de calificar MEJOR que otros segmentos";*

134. *Y a su vez, "Ciertos segmentos tuvieron la tendencia de calificar PEOR que otros segmentos";*

135. *"Referirse a ciertos aspectos hizo que las calificaciones se vean afectadas POSITIVAMENTE";*

136. O bien, "Referirse a ciertos aspectos hizo que las calificaciones se vean afectadas NEGATIVAMENTE";

137. "Las calificaciones que dieron 'MALO' están fuertemente asociadas con una clasificación del sentimiento NEGATIVA".

Pero esto no dejó pistas relevantes sobre las que podrían ser las variables relacionadas, o correlacionadas, o causales, de una calificación del tipo "bajo". Además, esta también se puede deducir de la primera proposición ("Las calificaciones que dieron 'BUENO' están fuertemente relacionadas con una clasificación del sentimiento POSITIVA"), sin necesidad de observar la regla correspondiente a esta. Por lo tanto, al ver que no se ha obtenido mucha información acerca de las variables que podrían afectar negativamente la calificación del usuario. Se optó por establecer la "clasificación por sentimiento" como variable consecuente, y así obtener pistas sobre las posibles causas o características que podrían afectar la clasificación por sentimiento del usuario. De esta forma, se obtuvieron las siguientes reglas[56]:

Figura 7.10. Lista de reglas generadas con la clasificación del sentimiento como variable consecuente

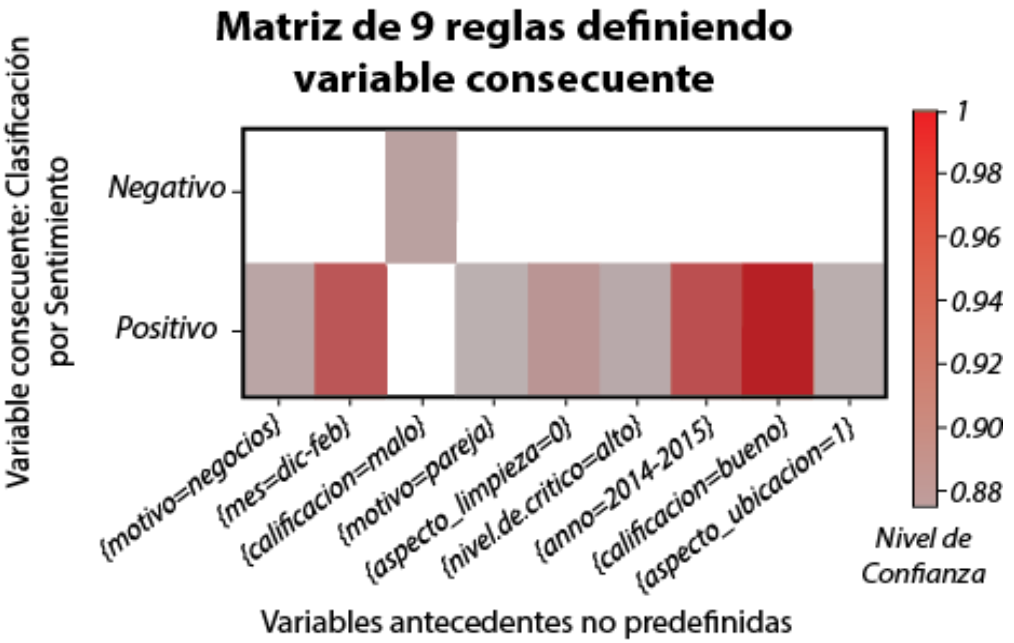
Antecedente		Consecuente	Soporte	Confianza	Lift
{calificacion=bueno}	=>	{sentimiento=positivo}	75%	100.00%	1.28
{anno=2014-2015}	=>	{sentimiento=positivo}	55%	94.83%	1.22
{mes=dic-feb}	=>	{sentimiento=positivo}	17%	94.44%	1.21
{aspecto_limpieza=0}	=>	{sentimiento=positivo}	33%	89.19%	1.14
{calificacion=malo}	=>	{sentimiento=negativo}	22%	88.00%	4.00
{motivo=negocios}	=>	{sentimiento=positivo}	14%	87.50%	1.12
{nivel.de.critico=alto}	=>	{sentimiento=positivo}	40%	86.96%	1.11
{aspecto_ubicacion=1}	=>	{sentimiento=positivo}	71%	86.59%	1.11
{motivo=pareja}	=>	{sentimiento=positivo}	19%	86.36%	1.11

Fuente: producción propia en base a datos de R.

A su vez, la siguiente imagen puede facilitar la visualización e interpretación de los datos:

[56] Se utilizaron los mismos criterios anteriores, con la excepción, de que esta vez el nivel de confianza se estableció con un mínimo del 85%.

Figura 7.11. Matriz de reglas generadas con la clasificación del sentimiento como variable consecuente



Fuente: producción propia en base a datos de R.

De este gráfico se podría haber deducido cuestiones casi idénticas al anterior (en donde se trata la calificación del usuario como variable consecuente) ya que las variables antecedentes son las mismas (a excepción del aspecto tecnología). Tampoco se pudieron obtener pistas relevantes para conocer las variables relacionadas, correlacionadas, o causantes de una clasificación por sentimiento del tipo "negativo"; por lo que a este capítulo respecta, no fue posible identificar las reglas de asociación en donde se involucra esta última variable y su categoría.

5.8. Capítulo 8: Estadística inferencial

5.8.1. Definición de criterios y de la técnica.

Según lo previsto en el marco teórico, realizar inferencias estadísticas significa asumir o afirmar ciertos atributos de la población a partir de una muestra; y además, se debe asumir un cierto nivel de riesgo a estar equivocado (Capriglioni, 2003b, p.13).

Para esta parte, se estableció la "calificación del usuario" como la variable principal a ser sometida a diversas pruebas. Sin embargo, al tratarse de datos que (como se vio en el capítulo 1: estadística descriptiva) parecerían no ajustarse a una distribución normal, y al no poseer un índice de asimetría aceptable, no se podrían hacer supuestos acerca de la población; y por lo tanto, las pruebas no podrían ser del tipo paramétricas. Además, se puso en duda la supuesta escala discreta de la variable principal, por más que esta vaya del 1 al 5, también TripAdvisor sugiere una escala paralela de 5 órdenes ("horrible" a "excelente"). Por lo que, suponiendo que es muy posible que esta variable es del tipo ordinal, correspondió que las pruebas realizadas hayan sido del tipo no paramétricas (Anderson, 2008, p.814).

Para estos datos, variables provenientes de Didi Soho Hotel y de Blue Soho Hotel, se contó con la herramienta R para aplicar las pruebas y obtener los resultados y gráficos de forma rápida. De esta forma, poder hacer hincapié en las inferencias y en la generación de información, en lugar de realizar un estudio exhaustivo sobre los aspectos estadísticos. Las pruebas fueron realizadas contemplando un nivel de confianza del 95%, debido a que la práctica hizo frecuente su utilización (Anderson, 2008, p.304).

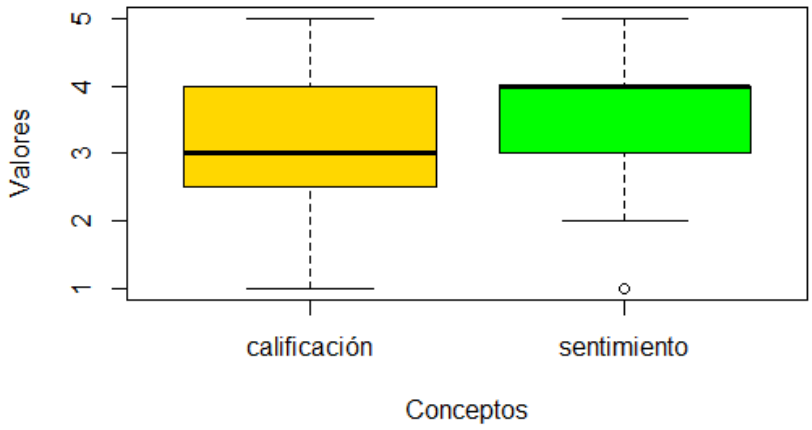
5.8.2. Prueba de rangos con signo de Wilcoxon.

La prueba de los rangos con signo de Wilcoxon es la alternativa no paramétrica al método paramétrico de las muestras por pares (o pareadas). En donde cada unidad experimental genera dos observaciones, correspondientes a dos poblaciones diferentes (Anderson, 2008).

En esta prueba se intentó observar las diferencias entre la variable principal o "calificación del usuario" y la "clasificación por análisis del sentimiento" transformada a escala numérica (ver capítulo 4: clasificación por análisis del sentimiento), y de esta forma, suponer si existen diferencias significativas entre estas dos variables pareadas. Antes de realizar la prueba, se

confeccionó el siguiente gráfico de diagrama de cajas[57], con la intención de poner a prueba ciertos atributos de los y luego compararselos con la prueba realizada:

Figura 8.1. Diagrama de cajas de la calificación y la clasificación del sentimiento para la prueba de rangos con signo de Wilcoxon



Fuente: producción propia en base a datos que arroja R.

Dado el diagrama de cajas, para la calificación del usuario, se pudo detectar que la mediana es para la tiene un valor de 3 (en contraste con lo estudiado en el capítulo 1, en donde la mediana poblacional tenía un valor de 4) y su posición dentro del rango intercuartílico dice que existe asimetría a la derecha, por lo que los valores están más concentrados a la izquierda. Por otro lado, la mediana para la clasificación por análisis del sentimiento tiene una mediana de valor 4 (cuatro), y su posición dentro del rango intercuartílico dice que existe asimetría a la izquierda, por lo que los datos están concentrados a la derecha. Además, la posición de la mediana dice que para el cuartil 2 y 3, las medianas coinciden. Visto que el valor 1 (uno) es considerado un valor atípico, debería suponerse que, además de que el valor 1 está muy distante de los otros valores, o bien, no pertenece al 90% de la muestra. Dicho esto, podría apoyarse aún más la decisión de no hacer supuestos acerca de la población.

Para proceder con la prueba de rangos con signo, se estableció lo siguiente antes de realizar la prueba, en materia de estadística inferencial:

[57] El diagrama de cajas es un tipo de gráfico que suministra información acerca de la mediana, sobre el 50% y 90% de los datos, y los valores atípicos. Utiliza los valores máximos y mínimos como límites superiores e inferiores. Construye la caja con los cuartiles 1 y 3, conteniendo el 50% de los datos, y también con el cuartil 2 como la mediana. Si la línea de la mediana no está en el centro de la caja, podría suponerse que existe asimetría en la distribución. Los valores atípicos son representados con pequeños círculos.

Figura 8.2. Definición de elementos de estadística inferencial para la prueba de rangos con signo de Wilcoxon

Hipótesis de prueba: • H_0: no existen diferencias significativas entre las poblaciones • H_a: existen diferencias significativas entre las poblaciones
Se quiere probar que la diferencia entre medianas es = 0
Nivel de confianza = 95%
$\alpha = 0.05$
Prueba con dos colas

Fuente: tabla de elaboración propia

Mediante la utilización de R[58], se obtuvieron los siguientes resultados:

Figura 8.3. Resultados de la prueba de rangos con signo de Wilcoxon

Suma de los rangos con signo = 35
Regla de decisión: • Si p-valor < 0.025 ó > 0.975: se rechaza H_0 • Si p-valor > 0.025 y < 0.975: NO se rechaza H_0
p-valor = 1.077^{-9}
p-valor es menor a 0.025. Se rechaza H_0, por lo tanto: • H_a: existen diferencias significativas entre las poblaciones

Fuente: tabla creada por el investigador con los resultados de R

Visto que la hipótesis alternativa mostró que tanto las calificaciones del usuario y el resultado del análisis del sentimiento presentan diferencias, es posible formular lo siguiente:

138. "Ambas variables miden dos cosas diferentes, por lo que podrían ser de utilidad para cualquier estudio a futuro."

Si bien se desconocen los intervalos entre las órdenes de ambas variables en cuestión, estos intervalos no son los mismos para cada variable, es decir, que por ejemplo para el análisis del sentimiento, se desconoce el espacio que hay entre "N+" y "N", entre "N" y "NEU", entre "NEU" y "P", y entre "P" y "P+". Lo mismo para la calificación del usuario en donde la escala va de 1 a 5, y paralelamente de "horrible" a "excelente". Por lo que podría decirse con mayor seguridad que al iniciar este capítulo que:

[58] Paquete utilizado: "BSDA".

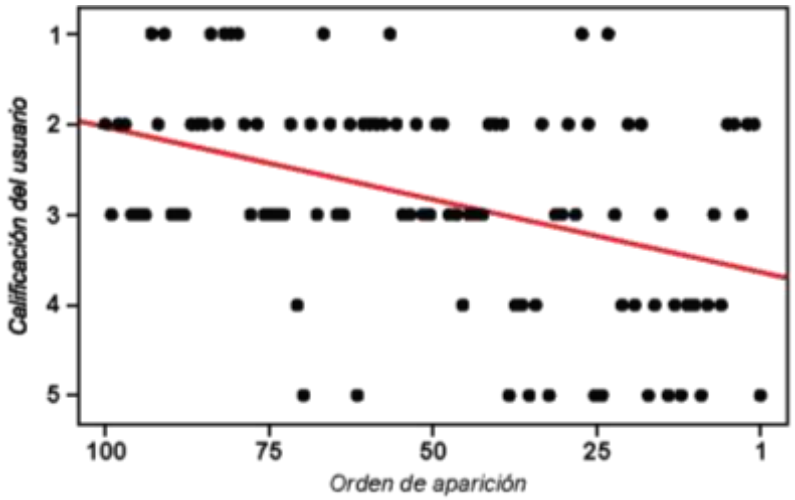
139. "Tanto la calificación del usuario como la clasificación por análisis del sentimiento no son variables en escala discreta, pero los intervalos de los valores son complejos de obtener."

5.8.3. Coeficiente de correlación de Spearman.

El coeficiente de correlación por rangos de Spearman es una medida de la relación lineal entre dos variables cuyos datos se encuentran en una escala ordinal (Anderson, 2008, p.837).

Para este caso, se intentó observar si podría existir alguna tendencia acerca de las "calificaciones de los usuarios" y el "orden de aparición" en TripAdvisor. Por lo que, la variable explicativa fue el orden de aparición y la variable explicada fue la calificación, en donde la variación de la variable explicada podría estar sincronizada con la variación de la explicativa. De esta forma, mediante R[59], se obtuvo que el índice de correlación de Spearman fue de: $\rho = -0.383827$. Por lo que, se logró establecer que existe una relación negativa entre ambos rangos; es decir, que a medida que fueron apareciendo más opiniones, la tendencia de las calificaciones fue decayendo. El siguiente gráfico sostiene dicha tendencia:

Figura 8.4. Diagrama de correlación de Spearman



Fuente: producción propia en base a datos que arroja R.

Esta prueba solo complementó con otro tipo de información lo expuesto anteriormente acerca de la tendencia de las calificaciones del usuario a lo largo del tiempo (ver capítulo 7: reglas de asociación). Basándose en esta tendencia y en los resultados vistos a lo largo del trabajo:

[59] Ningún paquete fue necesario para esta prueba.

140. "Esta tendencia negativa en las calificaciones persistirá a medida que aparezcan más opiniones en TripAdvisor si no se toman decisiones tendientes a mejorar las calificaciones del usuario."

5.8.4. Prueba de aleatoriedad.

La prueba de aleatoriedad o prueba de rachas comprueba si los datos se han extraído de forma aleatoria. Por lo que, para poder realizar esta prueba, se deben conservar los datos por el orden de observación (Cáceres, 1995, p.322). Esta prueba precisa que la variable sea dicotómica. Las rachas suceden una vez que hay un cambio en el valor de la variable.

Para este caso, también se utilizó la variable principal, calificación de los usuarios. Se utilizó R[60] para llevar a cabo el experimento, que por medio de su propio algoritmo, transformó la variable en cuestión a una dicotómica de la siguiente forma:

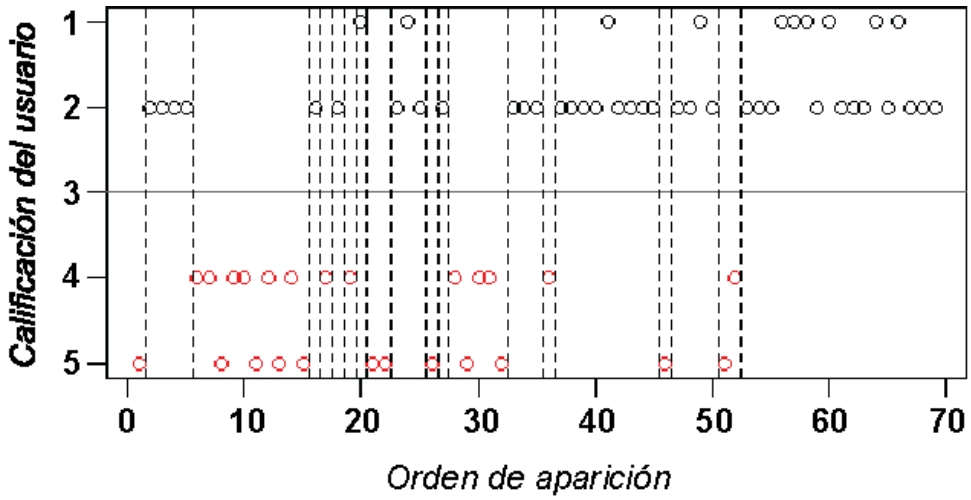
Figura 8.5. Escala de transformación de la calificación para la prueba de Aleatoriedad

Calificación del usuario (original)	Valores "1" y "2"	Valores "3"	Valores "4" y "5"
Calificación del usuario (dicotómica)	0	Excluido del experimento	1

Fuente: tabla creada por el investigador de acuerdo con lo estipulado por R

Esto se pudo representar en el siguiente gráfico:

Figura 8.6. Gráfico de representación para la prueba de Aleatoriedad



Fuente: producción propia en base a datos que arroja R.

[60] Paquete utilizado: "randtests".

Los criterios para el experimento fueron los siguientes:

Figura 8.7. Definición de elementos de estadística inferencial para la prueba de Aleatoriedad

Hipótesis de prueba: • H_0: la muestra fue tomada de forma aleatoria • H_a: la muestra NO fue tomada de forma aleatoria
Nivel de confianza = 95%
$\alpha = 0.05$
Prueba con dos colas

Fuente: tabla creada por el investigador

R arrojó los siguientes resultados:

Figura 8.8. Resultados de la prueba de Aleatoriedad

Variable estandarizada (z) = -3.3857
Cantidad de rachas = 20; Cantidad de unidades (variable dicotómica) = 69; Cantidad de unidades "positivas" = 44; Cantidad de unidades "negativas" = 25.
Regla de decisión: • Si p-valor < 0.025 ó > 0.975: se rechaza H_0 • Si p-valor > 0.025 y < 0.975: NO se rechaza H_0
p-valor = 7.099^{-4}
p-valor es menor a 0.025. Se rechaza H_0, por lo tanto: • H_a: la muestra NO fue tomada de forma aleatoria

Fuente: tabla creada por el investigador con los resultados de R

Aquí se mostró que, según la hipótesis alternativa, no existe aleatoriedad por parte de TripAdvisor, en el orden en el que muestra las opiniones. Por lo tanto, podría decirse que TripAdvisor posee alguna tendencia a priorizar la fecha de viaje del consumidor. Pero esto parecería ser una ventaja, o bien una oportunidad, ya que, si bien la tendencia indica que las calificaciones vienen decayendo, es posible lograr un cambio en donde los usuarios califiquen más alto y estos sean los que aparezcan primero en TripAdvisor, por ser más actualizados, y aquellas calificaciones que sean malas y menos actualizadas queden al final de la lista.

También, se vio que existen 69 unidades en total, de las cuales 44 son positivas y 25 son negativas; de esta forma, se sabe que la calificación global fue más alta en épocas anteriores que en épocas recientes.

Por otro lado, R ignoró los valores "3" por considerarlos neutrales, o bien, el valor menos extremo. Pero qué pasaría si no se considerara el "3" como un valor neutral, es decir, que su

valor sea, por ejemplo, positivo. En este caso, R no debería hacer un corte entre el "2" y el "4", sino que debería hacerlo entre el "2" y el "3". Por lo que, apoyándose en lo dicho en las otras pruebas no paramétricas, esto sugirió con mayor seguridad que:

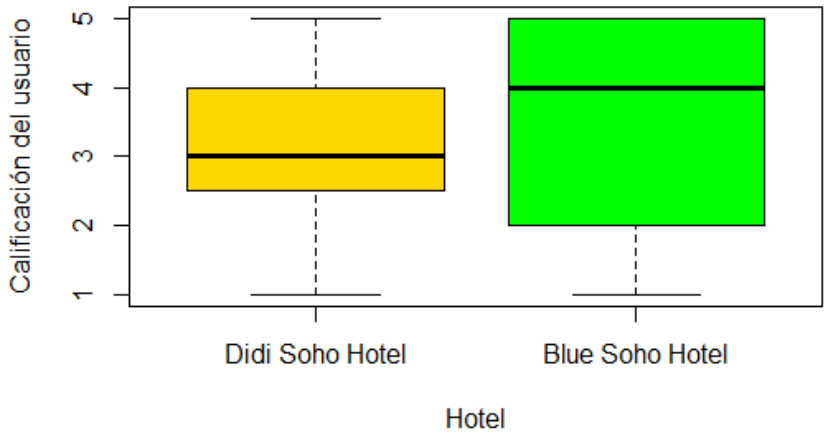
141. "La calificación del usuario no es una variable numérica discreta, sino que es una variable en escala ordinal".

5.8.5. Prueba U de Mann-Whitney.

La prueba U de Mann-Whitney es un método no paramétrico que se usa para determinar si hay diferencia entre dos poblaciones independientes, en donde las muestras no se pueden ordenar de a pares (Anderson, 2008, p.825).

Para este caso se intentó realizar una comparación entre las muestras de la variable principal para los dos hoteles tratados en este trabajo: Didi Soho Hotel y Blue Soho Hotel. A fin de realizar inferencias acerca de estas dos poblaciones. Previo a realizar la prueba, se confeccionó el siguiente gráfico de diagrama de cajas, solo para ver si a simple vista los datos pueden ser comparados prueba realizada:

Figura 8.9. Diagrama de cajas de los hoteles para la prueba de U de Mann-Whitney



Fuente: producción propia en base a datos que arroja R.

El diagrama de cajas muestra que, al igual que en la prueba de suma de rangos con signo de Wilcoxon, la calificación de Didi Soho Hotel posee una mediana de 3 y distribución asimétrica. En el caso de la calificación de Blue Soho Hotel, posee una mediana de 4 (coincidiendo con la mediana poblacional estudiada en el capítulo 1) y su distribución también es asimétrica hacia la izquierda, en donde los datos están concentrados a la derecha de la distribución. Además, tanto el límite superior de la caja (el cuartil 3) como el valor máximo coindicen con un valor de 5. No se registraron valores atípicos. Dicho esto, podría decirse que:

142. "Existen diferencias significativas entre las muestras de los dos hoteles."

Dicho esto, lo siguiente fue establecido antes de realizar la prueba, en materia de estadística inferencial:

Figura 8.10. Definición de elementos de estadística inferencial para la prueba de U de Mann-Whitney

Hipótesis de prueba: • H_0: no existen diferencias significativas entre las poblaciones • H_a: existen diferencias significativas entre las poblaciones
Se quiere probar que la diferencia entre medianas es = 0
Nivel de confianza = 95%
$\alpha = 0.05$
Prueba con una cola

Fuente: tabla creada por el investigador

Mediante la utilización de R[61], se obtuvieron los siguientes resultados:

Figura 8.11. Resultados de la prueba de U de Mann-Whitney

Suma de los rangos con signo = 5687.5
Regla de decisión: • Si p-valor < 0.05: se rechaza H_0
p-valor = 0.08418
p-valor es mayor a 0.05. Por lo tanto, NO se rechaza H_0, por lo tanto: • H_0: NO existen diferencias significativas entre las medianas de las poblaciones

Fuente: tabla creada por el investigador con los resultados de R

Según lo indicado por esta prueba, se puede inferir que:

143. "las poblaciones de ambos hoteles son similares."

Es decir, que como competidores ninguno de los dos está a la cabeza. Sin embargo, considerando que el nivel de error es solo del 5%, si se hubiera tomado un nivel de error del 10%, el p-valor rechazaría la hipótesis nula. Por lo que, se podría llegar a dudar de la inferencia realizada por esta prueba.

[61] Ningún paquete fue necesario para esta prueba.

CONCLUSIONES Y FUTURAS CONTRIBUCIONES

6.1. Capítulo 1: Estadística descriptiva

En esta primera herramienta, se vio cómo las herramientas correspondientes de estadística descriptiva pueden brindar información no sólo en materia estadística y del hotel principal (Didi Soho Hotel). Sino que también sirve para generar información para fines estratégicos, ya que permite conocer a los competidores del mercado (Blue Soho Hotel y otros hoteles cercanos). Esta permitió conocer cómo se comportan las calificaciones de los turistas, la época del año en la que se hospedaron y el motivo de viaje; clasificables además según la lengua de uso cotidiano de los huéspedes (español, inglés o portugués).

De esta forma, mediante la comparación de los datos procesados por herramientas de estadística descriptiva, se pudo obtener una visión más clara sobre la imagen en línea de los dos hoteles según TripAdvisor. Luego, podría decidirse qué tipo de movida estratégica podría emplearse a futuro. A modo de ejemplo, se propone lo siguiente:

- Se vio que el lusoparlante por lo general percibe mejor el producto hotelero de Didi Soho Hotel por sobre el de Blue Soho Hotel, y los viajeros lusoparlantes cuyo motivo fue el viaje entre amigos se mostró como el segmento mayoritario en este hotel. Dicho esto, podría pensarse en alguna decisión estratégica tendiente a acaparar el segmento de turistas lusoparlantes que viajen entre amigos. Así, identificando este espacio del mercado no aprovechado, podría dar lugar a una oportunidad de mercado, de acuerdo con la estrategia del "Océano Azul" (Kim y Mauborne, 2005, p. 5).
- También se vio que los turistas angloparlantes de ambos hoteles fueron en su mayoría viajeros en pareja. Los cuales se presentaron en mayor cantidad entre los meses de diciembre y febrero. En el caso de considerarse este segmento como atractivo, podría pensarse en alguna decisión estratégica para acapararlo. Dicho esto, podría implementarse una estrategia genérica (precio, diferenciación, foro) o una movida competitiva (cooperativa, represalia, defensiva) (Porter, 1985, p. 34).

Restaría utilizar las mismas herramientas de estadística descriptiva sobre otros hoteles, los cuales en un principio serían los hoteles de gama media que se encuentren en Palermo Soho (Koten Hotel, Five Cool Rooms, Hotel Costa Rica, etc.). Luego el estudio seguiría con el resto de los hoteles de Palermo, o bien, con los hoteles de otra categorización o nivel de precios que no sea gama media. De esta forma, habría una mayor riqueza en cuanto a la calidad y cantidad

de la información generada. Además, incluyendo más hoteles, quizás la corroboración de las hipótesis creadas sea menos necesaria, ya que sería más difícil que terminen siendo observaciones erróneas.

Finalmente, otro tipo de técnicas y gráficos de estadística descriptiva, como aquellas para cantidades y frecuencias de relativas (Capriglioni, 2003a, p. 23), o bien, mediante técnicas para variables cuantitativas discretas (Capriglioni, 2003a, p. 43) considerando que la calificación del usuario está en una escala discreta.

6.2. Capítulo 2: Frecuencia de términos

En esta segunda técnica de minería de datos se vio una primera forma de resumir los datos de las opiniones en forma de lenguaje natural. Se vio cuáles son los términos más utilizados por los huéspedes al detallar sus experiencias. Esto sirvió para dos tipos de datos principales:

1) Primero, aquellas palabras que más definen a priori el hotel: "bueno", "excelente", "malo", entre otros;
2) Segundo, qué atributos o aspectos del hotel fueron mayormente señalados: "habitación", "ubicación", "atención", entre otros.

Estos quizás no hayan revelado ningún patrón de comportamiento o información valiosa y mínimamente confiable que pueda facilitar la toma de decisiones como se vio con la técnica anterior. Apenas se pudo sugerir posibles ventajas competitivas de cada hotel, siendo las habitaciones la de Didi Soho Hotel, y la atención o el personal para Blue Soho Hotel. Sin embargo, esta técnica fue necesaria no solo para obtener una primera mirada sobre el hotel, sino también este paso es un requisito indispensable para proseguir con el resto de las técnicas que ofrece la minería de textos y de opiniones.

Por otro lado, podrían mejorarse las trasformaciones de los términos (correcciones y modificaciones), con la finalidad de simplificar el número de términos, o bien, obtener una visión más refinada acerca del hotel.

Finalmente, al igual que en el punto anterior, este mismo estudio podría extenderse a otros hoteles de la misma gama y zona, o bien, fuera de estos parámetros.

6.3. Capítulo 3: Asociación de términos

Mediante esta tercera técnica de minería de datos se ha mostrado una forma de relación entre dos términos, estableciendo previamente si tener en cuenta o no algunos de los términos. Luego de imprimir los gráficos correspondientes, se construyeron las correspondientes proposiciones que puedan reflejar y resumir lo que posiblemente se haya dejado plasmado en las opiniones.

De esta forma, se pudo decir que, por ejemplo, las habitaciones son mayormente percibidas como amplias y cómodas, o que la ubicación es excelente, o que el ruido de los boliches y bares de la zona dificulta la calidad de descanso. Estas proposiciones reflejan una mirada propia y exclusiva del hotel, en donde se expone solamente cómo se percibió y describió los aspectos principales y secundarios del producto hotelero. En caso de que se lo desee, se pueden ver otros aspectos con solamente indicar el término que corresponde (“desayuno”, “restaurante”, “piscina”, “bar”, “gimnasio”, etc.). Difícilmente en este capítulo se pudieron concebir proposiciones generalizadoras que brinden información por fuera de la empresa hotelera (salvo los términos que se relacionan con el término “hotel”, los cuales no se han tenido en cuenta).

Así, esta técnica intenta reducir el nivel de incertidumbre del universo, y por lo tanto, facilitar la toma de decisiones. En donde, por ejemplo, si se obtiene una mirada más a fondo acerca de los aspectos principales del producto turístico como la ubicación del hotel y la calidad de las habitaciones, podría pensarse en formas de promocionar este producto como diferenciado teniendo en cuenta sus mejores atributos (Kotler, 2011, p. 103).

Por otro lado, ciertas cuestiones con respecto al PLN han dificultado el refinamiento de la técnica, como lo fueron:

- La imposibilidad de separar las opiniones en sub-oraciones (Pang y Lee, 2008, p. 29) [i], ya que las comas (“,”) no permiten distinguir entre el nombramiento de elementos (ejemplo: “las habitaciones son amplias, cómodas y seguras”) y cuando efectivamente se tratan de suboraciones en donde existen dos sujetos diferentes (ejemplo: “las habitaciones son cómodas, y las camas son grandes”);
- La imposibilidad de clasificar “términos de opinión” (Bing, 2012, p. 12) [i], en donde, por ejemplo, la palabra “cómodo”, podría emplearse de las siguientes formas:
  - “Las habitaciones son cómodas”;
  - “El personal nos hizo sentir cómodos”;
  - “Uno puede ejercitarse cómodamente en el gimnasio”.

6.4. Capítulo 4: Clasificación por análisis del sentimiento.

Esta técnica sirvió para observar dos cuestiones principales:

1) Simplifica la satisfacción general de los huéspedes en pocas palabras y/o en valores numéricos;
2) Señala cuáles son los términos cuya utilización podría estar relacionado con una baja o alta calificación.

En el punto 2), la información generada podría utilizarse complementariamente con la técnica del capítulo 3 (asociación de términos), para obtener una mejor visión sobre un aspecto (o sub-aspecto) y así poder optar por mejóralo o no. Por ejemplo, la imagen proporcionada por R (discriminando términos) muestra qué tan negativos o positivos fueron algunos de los aspectos y subaspectos. Por lo que debería evaluarse la posibilidad de mejorar aquellos que más impacten negativamente teniendo en cuenta la capacidad de inversión del hotel. Ya que, de esta forma, se estaría contribuyendo en mayor proporción a la obtención de mejores calificaciones de los próximos usuarios, y por lo tanto, mejorando la imagen del hotel en medios digitales.

Además, se vio que uno de los términos que mayor negatividad expuso fue "ruido". Pero este ruido, como se vio en el capítulo 3, proviene de boliches y bares de la zona en donde el hotel está ubicado, por lo que resultaría poco viable intentar mejorar la ubicación del hotel, ya que el costo asumido sería bastante alto. Por otro lado, en el caso del baño, este también expuso una negatividad considerable, pero esta podría ser mejorada a un costo financiero y a un plazo relativamente corto, y además, esta mejora de este atributo "baño", podría mejorar potencialmente la satisfacción del huésped, y por lo tanto, podría también mejorar la calificación del hotel en la plataforma.

Además, el punto 1) puede servir como variable de control para el punto 2), ya que ofrece valores numéricos y textuales que indican si efectivamente la calificación global estuvo por encima o por debajo de las demás calificaciones históricas. Por ejemplo, si se toman nuevamente muestras de 100 unidades en un futuro cercano, podrían pasar dos cosas con respecto a las Mo(x):

a. Que estas sean distintas, es decir, que por ejemplo de pasar a tener una Mo(x) con valor "P", a pasar a una Mo(x) con valor "P+";
b. O bien, que las Mo(x) sean las mismas en ambos casos, pero las frecuencias de estas Mo(x) hayan cambiado. Por ejemplo, habiendo tenido una Mo(x) con valor "P" con una frecuencia de 54, se pasó a tener una Mo(x) con el mismo valor "P" pero con una

frecuencia de 37. Para este caso habría que analizar que pasó con las frecuencias de los otros valores también, y viendo estos y comparándolos con la otra muestra, revisar para qué valores las frecuencias aumentaron o disminuyeron.

Mediante el add-in proporcionado por Excel, se podría decir que cumple con su objetivo en cuando a la clasificación por análisis del sentimiento. Sin embargo, al desconocerse mediante qué mecanismos el add-in clasifica las opiniones y qué tan certeramente lo hace. Sería lógico pensar que se necesitaría desarrollar un algoritmo con la herramienta R, siendo este propio, controlable y modificable para realizar las tareas de análisis del sentimiento. De esta forma, la herramienta podría ser adaptada a otros requerimientos que surjan a futuro.

Pero para ello, existen grandes obstáculos, los cuales deben ser considerados. Algunos de estos podrían tratarse de:

Al igual que en el capítulo 3, prevalece la imposibilidad de clasificar "términos de opinión" (Bing, 2012, p. 12) [i], en donde, para este caso, un mismo término podría ser empleada para expresar un sentimiento positivo o negativo, por ejemplo:

- "El precio por las habitaciones es barato" (positivo);
- "Nos sirvieron comida barata" (negativo);

O también, enunciados más ambiguos, como lo son:

- "Ese error nos costó dentro de todo barato" (positivo dentro de un negativo);
- "Terminó siendo barato" (se necesita información anterior para clasificarlo como positivo o negativo).

Por otro lado, en el caso de contar con dobles negaciones[62], por ejemplo:

- "Ningún problema!" (connotación positiva);
- "No fue del todo malo" (sugiere que hubo algo malo, pero podría ser neutro).

[62] No existe bibliografía al respecto que hable del valor semántico de las dobles negaciones del PLN en inglés, ya que dicho idioma no reconoce las dobles negaciones como gramaticalmente correctas.

Por otro lado, debería tenerse en cuenta cuando dos términos negativos, no implican positividad ni neutralidad, como lo son:

- "Feo y sucio" (dos términos negativos que potencian lo negativo);
- "No hospedarse jamás!" (doble negación en una sola oración).

Por otro lado, también existen ciertas cuestiones que muestran que existe un problema realmente desafiante (Bing, 2012, p. 52). [i], tales como:

- Interjecciones: "yeah!", "mmm…", "bueno…", etc.;
- Número de puntuaciones: "muy bueno" vs "muy bueno!" vs "muy bueno!!!";
- Número de caracteres en mayúsculas: "Muy bueno" vs "MUY BUENO";
- Emoticones: ":)", ":(", ";)", etc;
- Ironías y sarcasmos:
  - "Muy buena la ducha! Me morí de frío…" (palabras de connotación positiva utilizadas de forma irónica o sarcástica);
  - "Este "hotel" está "bien" ubicado" (insinuando mediante la utilización de comillas ("") que no se trataba de un hotel formal, y que tampoco estaba tan bien ubicado como se creía).
- Opiniones falsas o "spam":

6.5. Capítulo 5: Clasificación por subjetividad.

De acuerdo con la analogía propuesta entre la clasificación por subjetividad y el modelo de segmentación vincular, se mostró cómo esta técnica puede ayudar a exponer ciertos atributos de los segmentos mercado turístico que TripAdvisor define. Si bien, como se explicó anteriormente, la práctica de utilizar analogías conlleva a un alto riesgo de cometer un error en la decisión, la situación de extrema incertidumbre y ambigüedad, permitió su implementación (Bonatti et al, 2011, p. 76).

Gracias a la herramienta podría suponerse que algunos segmentos podrían estar más relacionados de acuerdo con su habla con un tipo de segmento en donde su vínculo con el producto hotelero es racionalista, es decir, estos segmentos son atraídos al producto hotelero más por su funcionalidad, relación precio-calidad, y practicidad, que por cuestiones afectivas, sentimentales o emocionales. Para esta segmentación racionalista, el segmento de viajeros que más se vio ajustada fue el de viajeros entre "amigos", seguidos por los que "familia", por lo

que se presume que un posicionamiento factible para estos segmentos sería por precio. Por otro lado, para los segmentos de negocios y pareja, dado que se tratan de segmentos que mostraron un mayor grado de subjetividad y un vocabulario más emocional, se indicó que podría ser preferible apostar a un posicionamiento por diferenciación por calidad del producto y/o una diferenciación por marca.

Una de las formas más simples para corroborar esta hipótesis de que los viajeros entre amigos sean los más racionalistas, sería crear una tabla pivot en Excel con las variables "motivo de viaje" y "aspecto calidad-precio", y ver si este tipo de viajero ocupa una proporción importante de aquellos quienes manifestaron interés sobre el precio del producto hotelero. De esta forma, seguir sosteniendo o no la hipótesis.

Por otro lado, también se podría realizar la misma clasificación por subjetividad con el resto de las opiniones del resto de los hoteles en Palermo Soho. De esta forma, seguir corroborando a un nivel por encima de un simple hotel, si existe algún segmento o segmentos que se ajusten más al tipo de vínculo "racionalista".

Al igual que como en este capítulo se realizaron analogías (cuyos riesgos han sido tenidos en cuenta), se podrían realizar otras analogías que podrían llegar a brindar información relevante acerca del mercado para la toma de decisiones, conllevando a estrategias competitivas y de marketing. Estos modelos pueden ser: la Cruz de Porter, la Matriz FODA, la Matriz BCG, la Matriz de Ansoff, la Matriz de Mckinsey, etc.

6.6. Capítulo 6: Generación de tópicos

Como se ha visto, la técnica de generación de tópicos brinda información acerca de las declaraciones de los huéspedes; particularmente, señala la cantidad de aspectos a la cual suele referirse un usuario en sus opiniones. Se ha asignado una alta probabilidad de que el huésped suela comentar entre 3 a 5 aspectos cada vez que contrata o consume un producto turístico. Por lo tanto, en cuanto a la gestión del hotel, sería lógico suponer que el hotel posee entre 3 y 5 oportunidades de impactar positivamente y satisfacer las demandas del consumidor; ya que, el huésped suele percibir esa misma cantidad de aspectos. Por otro lado, si esta oportunidad no es tomada en cuenta, esta podría generar un impacto no neutro, sino negativo.

Así, entonces, se expuso cómo esta técnica posee un cierto nivel utilidad para la gestión de recursos en un hotel. Los resultados de esta técnica podrían ser combinados con otras técnicas

anteriores como la asociación de términos y la clasificación por análisis del sentimiento (esta última, mediante la herramienta R). En donde habría que evaluar, como se mencionó en puntos anteriores, qué tanto un aspecto podría ayudar a mejorar el producto hotelero y la imagen del hotel en medios digitales. En la siguiente técnica (reglas de asociación) se expuso cuáles podrían ser los aspectos los cuales implicaron una mejor o peor calificación, por lo cual, esto podría dar una pista sobre qué aspectos podrían priorizarse para un mayor impacto en los huéspedes. De esta forma, poder contar con un modelo de decisión que permita mejorar el producto del hotel.

Con respecto a la supuesta normalidad de la cantidad de aspectos tratados por huésped, se previó que esta hipótesis prevalecerá por un buen tiempo para este hotel en particular. Lo que sí se podría intentar, es realizar un relevamiento por zona geográfica (barrio, ciudad, provincia, país, continente, etc.) para probar si realmente este es un patrón que existe y que pueda generalizarse acerca de los huéspedes.

En cuanto a las mejoras de la herramienta, debería considerarse lo siguiente:

1) El algoritmo utilizado simplemente detecta y compara las opiniones con un glosario de términos, los cuales son excluyentes para cada tópico. Es decir, la palabra "cómodo" solamente indicó que el tópico es "habitación"; pero qué pasaría si el usuario dice que el recepcionista del hotel lo hizo sentir "cómodo", en este caso, que podría ser menos frecuente que el anterior, el tópico sería calidad de "servicio" del personal.
2) Segundo, podrían sumarse otros tópicos, que no correspondieron al caso práctico presentado, como podrían serlo el "estacionamiento", "piscina", "restaurante", "gimnasio", "cancha de tenis", "excursiones", "actividades", etc. También, existen otros como el "mantenimiento" y el "baño", pero al estar muy entrelazados con la "limpieza" y las "habitaciones respectivamente, esto podría generar confusiones, pero aún así se los podría considerar.

Una forma de seguir corroborando estas hipótesis sería utilizando una mayor cantidad de tópicos, como se lo desarrolló en el párrafo anterior. De esta forma podría ocurrir lo siguiente:

1. Que el coeficiente de curtosis se aproxime a un valor 0, o bien, que termine siendo positiva, señalando un mayor empinamiento y por lo tanto una mayor concentración en los picos;

2. O bien, que este coeficiente se haga cada vez más negativo, el cual indicaría que cada vez más opiniones se salen del supuesto intervalo de 3 a 5 tópicos, al aumentar los valores por debajo del 3 y por encima del 5;
3. Por otro lado, en cuanto al coeficiente de asimetría, podría ocurrir que cada vez más opiniones se sitúen a la derecha del pico, en lugar de a la izquierda;
4. O bien, permanezca simétrico, es decir, que ni a la izquierda ni a la derecha exista una mayor concentración.

Por otro lado, también existen herramientas para corroborar si una población posee una distribución normal (o Gaussiana). Algunas de estas pruebas son: Kolmogorov-Smirnov, Anderson-Darling, Shapiro-Wilk, Chi-Cuadrada de Pearson, entre otros.

6.7. Capítulo 7: Reglas de asociación

Durante el transcurso de esta séptima parte, se pudo evidenciar cómo la generación de reglas de asociación puede ser utilizada tanto para generar información como para abrir nuevas líneas de investigación. Esta es una de las técnicas que más ampliamente puede ser utilizada, ya que permite procesar grandes cantidades de variables, siendo este el motivo por el cual este es el séptimo capítulo del total de 8 capítulos desarrollados en este trabajo. Además, se vio la gran cantidad de información que puede generar, en donde limitándola a solo un tipo de algoritmo y un solo tipo de medición, fácilmente supera los 3 millones de reglas, partiendo de tan solo 23 variables en su mayoría dicotómicas. En este capítulo, por cuestiones prácticas, se redujo esta cantidad a tan solo 302 reglas de acuerdo con los criterios mencionados.

Por otro lado, se vio que estas reglas poseen 3 tipos de mediciones (confianza, soporte y lift); y además, de estas reglas se pueden generar otras 4 utilizando las categorías de las variables opuestas a la original, o bien, en ambas variables. Por lo que, de estas 302 reglas, podrían deducirse (del producto de las mediciones y aquellas posibles reglas con lenguaje negativo) un total de 3.624 (que proviene del producto entre 302, 3 y 4) proposiciones como información o hipótesis, cuyas relevancias, significancias y utilidades sería juzgada en un trabajo a futuro.

Se consideró, además, que la cantidad de variables involucradas debieron ser 2 (una antecedente y otra consecuente), ya que de esta forma se facilita el entendimiento por parte del factor humano. Sin embargo, podría contemplarse el empleo de más de dos variables; esto es, adicionando más variables antecedentes.

Esta técnica permitió también definir cuáles variables involucrar. En este capítulo se trataron los segmentos del mercado hotelero, las calificaciones y clasificación del sentimiento. Pero esta focalización, si bien fue fructífera, no se lo utilizó a todo su potencial. Por lo que quedaría, entonces, para trabajos futuros poder ampliar la focalización hacia otras variables y sus categorías a fin de aumentar la posibilidad de obtener más información; y también, utilizar criterios menos limitantes, como lo son la reducción del nivel de confianza y el nivel de soporte a la hora de generar más reglas.

Por otro lado, mientras se tome una muestra con una mayor cantidad de opiniones (siempre y cuando se respete del criterio de "opiniones actualizadas"), posiblemente la reducción de categorías para cada una de las variables sea menos necesaria. En donde se podría contar con una mayor cantidad de categorías en total, y por lo tanto, se podrá generar una mayor cantidad de reglas de asociación que representen una información más refinada, o bien, con las mismas variables, obtener reglas con índices más certeros.

Por último, se logró mostrar una serie de variables que podrían ser determinantes para obtener una buena calificación o una clasificación del sentimiento positiva; sin embargo, no se logró hacer lo mismo con respecto a una mala calificación o a una clasificación del sentimiento negativa. Por lo que, en otros trabajos podría utilizarse un foco mayor hacia estas variables y sus categorías, ya sea reduciendo el nivel mínimo de confianza, soporte o lift, o bien, buscar otras técnicas de minería de datos que puedan facilitar su entendimiento.

6.8. Capítulo 8: Estadística inferencial

En esta última parte, se mostró cómo algunas de las pruebas de hipótesis pueden ser aplicadas tanto a los datos obtenidos de TripAdvisor como a otras variables generadas a lo largo del trabajo. Estas pruebas fueron del tipo no paramétricas, ya que el tipo de datos hizo que no sean aplicables las pruebas paramétricas. En donde el papel principal del investigador fue realizar inferencias que puedan ser utilizadas como información, o bien, para generar nuevos supuestos acerca de las opiniones.

En la prueba de suma de rangos con signo de Wilcoxon, se sugirió que la "calificación del usuario", no es una variable discreta, sino una variable ordinal (por más que vaya del 1 al 5). Con esta lógica, también podría suponerse la creación de una nueva variable que agregue los puntajes en uno final. En donde la variable principal "calificación del usuario", se vería modificada de acuerdo con la "clasificación del sentimiento", "clasificación por subjetividad",

"cantidad de tópicos", "nivel del crítico", "fecha de estadía", "orden de aparición", etc. De esta forma, obtener una variable continua, que pueda ser sometida a una mayor cantidad de pruebas no para métricas, o incluso paramétrica. Por lo que, a futuro, podría asignarse a esta variable, que resume a las otras variables, una distribución normal, binomial, t de student, exponencial o cualquier otro tipo de distribución conocida.

Acerca del índice de correlación de Spearman, se podrían realizar en trabajos futuros experimentos estadísticos con otras variables ordinales y someterlas a la prueba de coeficiente de correlación de Spearman (o Pearson en caso de poseer variables continuas). Algunos ejemplos podrían ser:

- "calificación del usuario" y "análisis del sentimiento": en donde se pondría a prueba lo que otros autores afirman acerca de la dependencia, correlación, e incluso causalidad entre estas dos;
- "análisis del sentimiento" y "orden de aparición": tal como se lo hizo en este punto, solamente que, en lugar de utilizar las calificaciones, se estaría viendo cómo evolucionó el sentimiento de acuerdo con como TripAdvisor lo muestra en su plataforma;
- "calificación del usuario" y/o" análisis del sentimiento" con la "cantidad de tópicos sin discriminar": en donde se podría ver si la cantidad de aspectos hacia los cuales un huésped incluye en su comentario tiene alguna relación con su sentimiento o calificación;
- "análisis del sentimiento" y los distintos coeficientes de objetividad (no como variable dicotómica): aquí también, según los teóricos, existe una relación particular entre estas variables, ya que una opinión objetiva haría que el sentimiento sea neutro; aún así, esto podría ponerse a prueba y ver si en la realidad sucede.

Acerca de la prueba de aleatoriedad, teniendo en cuenta las dudas sobre la escala de la variable calificación del usuario, se podría suponer que la diferencia entre un valor positivo y un valor negativo, no tendría que pasar por el valor del medio (una calificación de "3"), sino que, al tratarse esta de una variable ordinal, la diferencia entre negativo y positivo podría situarse entre el "2" y el "3". En donde "1" y "2" serían valores negativos, y los valores "3", "4" y "5" serían valores positivos.

Acerca de la prueba de Mann-Whitney se vio que la inferencia es un tanto dudosa, ya que viendo los datos en un gráfico de barras pareciera que, al menos las muestras, muestran

patrones diferentes. Por lo que, vista la poca cantidad de categorías de la variable, dispersar o incrementar la categoría de estas podría llegar a mejorar la exactitud de la prueba, ya que la etapa de la suma de rangos tendría resultados más dispersos, debido a una mayor cantidad de opciones. Incluso se vio que los diagramas de caja poseen este mismo inconveniente, ya que los rangos intercuartílicos, mediana, y valores atípicos son forzados a situarse en uno de estos cinco valores.

Por otro lado, es necesario aclarar que existen cantidades de otras pruebas no paramétricas que podrían ser aplicadas. Pero no correspondieron al presente trabajo por alguno de los siguientes: los datos tomados no son apropiados o no poseen consistencia, no corresponde al presente trabajo por el diseño de la investigación, o se necesitan más datos para ser implementadas. Como por ejemplo, si se contara con un grupo de usuarios que hayan estado en los dos hoteles (Didi Soho Hotel y Blue Soho Hotel) y ambos hayan dejado sus comentarios, los datos pareados podrían fácilmente ser sometidos a una prueba de rangos con signo de Wilcoxon (Anderson, 2008, p.820); y en el caso de contar con más hoteles, contando con un total de 3 o más, la prueba de Friedman (Cáceres, 1995, p.343). Además, tal como se realizó la prueba de U de Mann Whitney, si se contaran con muestras de otros hoteles, sencillamente podría aplicarse la prueba de Kruskal-Wallis (Anderson, 2008, p.833). Estos otros hoteles, según lo previsto por TripAdvisor y considerando las similitudes con los hoteles tratados en este trabajo, podrían ser: Five Cool Rooms, Hotel Bys Palermo, Koten Hotel, 1555 Malabia House, entre otros.

Para finalizar, se apreció que una mayor dispersión, o cantidad de valores se lograría, por ejemplo, condicionando los valores de la variable principal (calificación del usuario) de acuerdo con el valor del resultado de la clasificación por análisis del sentimiento, nivel de crítico, orden de aparición, cantidad de tópicos, etc. Esto debería hacerse, además, de acuerdo con una escala, la cual según la viable podría ser creciente o decreciente, o bien lineal, geométrica, o logarítmica. Futuros trabajos podrían encargarse de ello.

6.9. Consideraciones finales

A lo largo de este trabajo se ha logrado exhibir cómo algunas de las herramientas de la minería de datos pueden ser aplicadas sobre los datos provenientes de las reseñas que los usuarios dejan en TripAdvisor, y a través de ello, poder localizar información acerca del hotel y del turista, y además poder proponer nuevas líneas de investigación a ser realizadas a futuro, dentro y fuera de las disciplinas que se han involucrado.

Si bien asumir o reconocer la información que ha sido generada conlleva a asumir un cierto nivel de riesgo, debido al desconocimiento de la efectividad de las herramientas y falta de estudios pasados que demuestren la utilidad de estas herramientas, esto significó un punto de partida para explorar y explotar un área poco investigada. Intentando responder a las necesidades de información, tiempo y recursos informáticos para el decisor de una empresa turística (o cualquier otra organización de cualquier otro rubro).

Junto a esta información, se han propuesto diversas utilidades sobre cómo estas pueden ser implementadas para tomar una decisión. En ocasiones, también, cómo podría refinarse la información sea mediante mejoras en las técnicas, o bien, al tener en cuenta otro tipo de información proveniente o no de la misma fuente. Estas podrían abarcar los siguientes:

1. Muestras de otro tamaño: en el presente trabajo se estableció una muestra de 100 de forma arbitraria, por lo que parecería lógico si tomando un tamaño de muestra menor o mayor a 100 se podría asegurar un cierto nivel de representatividad en cuanto a tamaño y actualización de las opiniones. Si un hotel posee menos de 100 opiniones, se podrían tomar la totalidad de las opiniones, o bien una parte siempre y cuando se aseguren los criterios de representatividad. Sin embargo, si un hotel que posee miles de opiniones, debería investigarse si la muestra de 100 es suficiente o se requiere una muestra mayor para obtener información más refinada.
2. Otros datos del mismo hotel en TripAdvisor: si bien se supuso que ciertos datos no serían convenientes para la presente investigación, esto no quita que puedan ser útiles para otras investigaciones. Por lo tanto, los datos descartados como: fecha de publicación, calificaciones por aspecto, cantidad de fotos subidas, fotos subidas, contribuciones de los usuarios, recomendaciones de las habitaciones, etc., podrían ser de utilidad para obtener otro tipo de información, o bien, obtener información más refinada. Estos datos podrían ser:
3. Opiniones escritas en otros idiomas en TripAdvisor: como se explicó anteriormente las técnicas de PLN de la lengua inglesa están a la vanguardia del estado de arte. En el presente trabajo se logró realizar un primer paso para ajustarlas a la lengua española. Quedaría para en otra ocasión, ajustar las técnicas a la lengua portuguesa, la cual mostró una cantidad de opiniones considerable, y además, se prevé que la configuración natural de esta lengua, no dificulte considerablemente la adaptación de las técnicas. Por otro lado, probar si las opiniones escritas en diferentes idiomas pueden conformar una

misma muestra; aunque una de las hipótesis propuestas enunció que estas posiblemente provengan de poblaciones diferentes.

4. Opiniones de otros hoteles en TripAdvisor: se enunció que en Palermo Soho existen otros hoteles de la misma categoría (Five Cool Rooms, Hotel Costa Rica, Infinito Hotel, Koten Hotel, etc.), de los cuales se podría obtener información tanto genérica de la totalidad de hoteles competidores en la zona, o bien, información acerca de cada competidor individual. Además, también esto podría extenderse tanto geográficamente (según la zona, sub-barrio, barrio, ciudad, provincia, país, continente, etc.), como por la categorización del hotel (de 1 a 5 estrellas).
5. Opiniones de otras plataformas de opiniones turísticas: si bien se explicó el porqué de la elección de TripAdvisor como fuente principal de datos, también podrían contemplarse otras como Booking.com, Google maps, Expedia, viajeros.es, etc., con el fin de obtener otro tipo de información, o bien, conformar el mismo conjunto de muestra. Aunque la efectividad, consistencia y veracidad de las opiniones en estas plataformas deberían ser investigadas.
6. Otro tipo de fuentes: en este caso podrían ser la base de datos del hotel, registro de huéspedes, el libro de quejas, foros, y cualquier otra fuente de datos que se refiera al hotel específicamente. O bien, otras fuentes de datos que posibiliten la aplicación de herramientas estadísticas y de minería de datos (tanto los tratados en este trabajo como los que no fueron tratados), como podrían serlo las encuestas de ocupación hotelera, de turismo (receptivo, emisivo y doméstico), de viajes en hogares, tanto dentro del país en cuestión como fuera del país.

Con respecto a las hipótesis generadas, en ocasiones se han brindado pistas con la intención de corroborar las hipótesis, esto determinaría si seguir sosteniendo o rechazar estas hipótesis.

Por otro lado, otras disciplinas (tratadas y no tratadas en el presente trabajo) cuales podrían contribuir a un mayor entendimiento o interpretación de los datos, o bien, estas disciplinas podrían beneficiarse del presente trabajo, las cuales podrían ser:

1. La estadística: si bien se trataron elementos de la estadística descriptiva y estadística inferencial, no se generaron modelos de estadística predictiva, los cuales son un importante activo de la inteligencia de negocios (BI – Business intelligence). Este estudio podría iniciarse con la definición clásica de probabilidad, en donde la ocurrencia del suceso aleatorio es el cociente entre la cantidad de casos favorables y los casos

posibles (Capriglioni, 2003a, p. 172). Finalmente, adaptar si es posible las poblaciones a una distribución normal, y aplicar herramientas estadísticas específicas para las poblaciones normales, siendo que estas son por lo general más potentes que las pruebas no paramétricas (Anderson, 2008, p. 813);

2. La minería de datos: técnicas no tratadas durante el trabajo que procesan datos puramente numéricos, y sobre todo, es de suma importancia mejorar las técnicas existentes u otras no aplicadas en este trabajo de PLN. Ya que estas, podrían ser utilizadas, o bien, mejorando las anteriores, con la finalidad de obtener información más certera, o también mayor cantidad de información. Algunos de estos son: la detección de ironías y sarcasmos (Bing, 2012, p. 52), detección de opiniones "spam" (Bing, 2012, p. 14), resumen de opiniones (Bing, 2012, p. 102), entre muchos otros.
3. La lingüística y las ciencias médicas: se trata de herramientas de las disciplinas que estudian el habla de las personas, ya que por un lado, se necesitaría un mayor refinamiento del PLN. Además, otras ramas como la neurolingüística, la psicolingüística, la etnolingüística y la dialectología hispánica (si solo queremos estudiar las opiniones en español), podrían contribuir con la generación de información brindando una mirada más profunda sobre los consumidores. Resolviendo, entre muchas otras, ciertos problemas como los siguientes:
    a. El valor semántico de las palabras que expresan opinión, en donde decir que la ubicación del hotel es "excelente", quizás sea mejor a que uno dijera que la ubicación del hotel es "buena", y a su vez estas dos sean mejor que decir que la ubicación del hotel es "mala". Pero ¿Cuál sería el valor para cada una de estas palabras de opinión? Determinar el peso de cada término sería de gran valor para mejorar las técnicas de análisis del sentimiento;
    b. El valor semántico de una palabra equivalente según su traducción literal, en donde oraciones como "el hotel está bien ubicado" (en español), "the hotel is well located" (en inglés) y "o hotel está bem localizado" (en portugués), no deberían tener el mismo valor semántico, aunque su traducción literal sea esa. Ya que como se ha visto en el capítulo 1: estadística descriptiva, las poblaciones separadas según su idioma apreciaron o percibieron el mismo producto hotelero de forma diferente, por lo tanto estas poblaciones no deberían ser unificadas mediante la traducción;
    c. Lo mismo podría ocurrir dentro de un mismo idioma, en donde, por ejemplo, el hecho de que el hotel tenga una buena ubicación no tenga el mismo peso para un turista que proviene de la Ciudad Autónoma de Buenos Aires que para un sujeto

que provenga del sur de la Argentina. Podría ocurrir o no, que para el porteño, una buena ubicación forma parte de sus expectativas básicas, pero para alguien de un pueblo o zona muy poco poblada, la ubicación y cercanía a otras atracciones sea un punto mucho más favorable que para el porteño. Por lo tanto, sería lógico pensar que esto impactaría al puntaje general y a la clasificación por análisis del sentimiento de una forma positiva.

   d. Acerca de la clasificación por subjetividad: seguramente existan diferencias entre los distintos idiomas y dialectos, e incluso según la zona geográfica de donde provenga el huésped. Por lo que, la lingüística y sus ramas podrían ayudar a mejorar la tarea de clasificación por subjetividad, y de esta forma, determinar si es posible aplicar un modelo de integración con las opiniones y así podría ayudar a determinar si es conveniente su utilización para clasificar la categoría de un hotel (como se vio en las conclusiones del capítulo 5: clasificación por subjetividad).

4. El marketing y la competitividad: se vio cómo durante el transcurso del trabajo, cierto tipo de información da lugar a la ejecución de ciertas estrategias de marketing y movidas estratégicas de competitividad. Referido a estos, existe un campo inmenso y diversos autores los cuales sugieren toda clase de decisiones tendientes a la mejora de la rentabilidad de la empresa. Algunos de estos autores tratados en el trabajo fueron: Porter, Kotler, Levy, Wilensky, y otros no tratados como Hax, Meerman, Berger, Magretta, entre muchos otros;
5. Teoría de la decisión y de la información: para este trabajo se asumió que el universo se presenta en un nivel de ambigüedad (Bonatti et al, 2011, p. 76). Sin embargo ¿se podría considerar en otra categoría de incertidumbre una vez que se obtenga un mayor grado de conocimiento, como lo son el nivel de "incertidumbre", "riesgo" y "certeza", junto a sus respectivas herramientas (Bonatti et al, 2011, p. 76)? O bien, ¿proponer analogías y modelos dinámicos no lineales (Bonatti et al, 2011, p. 94)? O también, ¿aplicar conocimientos de la Física, como lo es la Teoría del Caos desarrollada por Ilya Prigogine, en donde se prevé que existen estructuras que mantienen un orden y escenarios impredecibles (Bonatti et al, 2011, p. 51) ?;
6. Turismo en los medios digitales: visto que en TripAdvisor existen más categorías además de los hoteles, tales como: "destinos", "aerolíneas", "alojamientos" y "actividades", las técnicas utilizadas podrían adaptarse a estas categorías, por lo que el conocimiento del turismo como disciplina podría ayudar a desarrollar nuevamente las técnicas implementadas para otras secciones que ofrece TripAdvisor. Por otro lado,

obtener una mirada de estas también contribuiría con una mejor planificación de un espacio turístico, en donde se podría obtener información valiosa acerca de ciertos elementos del sistema turístico, como lo son: la oferta turística, la demanda turística, el producto turístico y la planta turística (Boullon, 2006, p. 32).

7. En materia legal: algunos países[63] poseen un sistema completo o parcialmente integrado que utiliza las opiniones en medios digitales para definir la categorización de los hoteles (UNWTO, 2014, p. 17) [i]. ¿Podría entonces pensarse en tomar las opiniones en plataformas online para categorizar los hoteles en Argentina o en Latino América? De esta misma forma, otras entidades públicas, como por ejemplo la AFIP (Administración Federal de Ingresos Públicos), ente de ejecución de la política tributaria y aduanera de la Nación Argentina (Decreto Nº 618/1997), podría utilizar estas fuentes públicas de datos para detectar si existen inconsistencias entre la cantidad de opiniones y las fechas de las mismas, y las declaraciones por parte del hotel correspondiente; en otras palabras, fuera de las disciplinas involucradas y fuera de las entidades con fines de lucro, las opiniones actuarían como pistas de evasión impositiva.

Dicho esto, es posible afirmar que todavía existe un espacio enorme representado por el desconocimiento del tema tratado, y por lo tanto, esto merita estudios futuros que sirvan de aporte para comprender mejor esta pequeña porción el universo turístico.

[63] Estos países son Emiratos Árabes Unidos, Noruega, Suiza, Alemania y Australia.

FUENTES BIBLIOGRÁFICAS Y FUENTES DE DATOS

7.1. Textos, tesis y artículos

Anderson, D., Sweeney, D. y Williams, T. (2008). *Estadística para administración y economía.* (10ma Ed.). México D.F.: Cengage Learning Editores, S.A.

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. (2005). Análisis de una situación de decisión del mundo de la estrategia, con altos niveles de incertación y complejidad: la ocupación de las Islas Malvinas (Tesis de doctorado). Facultad de Ciencias Económicas, Universidad de Buenos Aires, Ciudad Autónoma de Buenos Aires, Argentina.

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.

Boullon, R. (2006). Planificación del Espacio Turístico. (4ta Ed.). México: Editorial Trillas.

Capriglioni, C. (2003a). *Estadística. Tomo I.* Argentina, Ciudad Autónoma de Buenos Aires: 3C Editores.

Capriglioni, C. (2003b). *Estadística. Tomo II.* (3ra Ed.). Argentina, Ciudad Autónoma de Buenos Aires: 3C Editores.

Cáceres, R. (1995). *Estadística multivariante y no paramétrica con SPSS*. Aplicación a las ciencias de la salud. Madrid, España: Ediciones Diaz de Santos, S.A.

Castillo, M. y Panosso, A. (2011). Implicaciones epistemológicas en la investigación turística. Estudios y Perspectivas en Turismo, Vol. 20, 384-403. Sao Paulo, Brasil. Recuperado de: http://www.scielo.org.ar/pdf/eypt/v20n2/v20n2a07.pdf

Conde, E. y Amaya, C. (2007). El Producto Hotelero: visto como un conjunto de atributos tangibles e intangibles. *Gestión Turística*. Vol. 8(5), 75 – 83. Recuperado de http://4www.redalyc.org/articulo.oa?id=223314983006

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

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.

Hernández, R., Fernández, C., Baptista, P. (2014). *Metodología de la Investigación*. (6ta Ed.). México D.F.: McGraw-Hill.

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

Jafari, J. (2005). El turismo como disciplina científica. *Política y Sociedad*. Vol. 42(1), 39-56. Recuperado de: https://dialnet.unirioja.es/servlet/articulo?codigo=1307535

Kim, W. C. y Mauborne, R. (2005). *Estrategia del Océano Azul*. España, Cataluña, Barcelona: Grupo Editorial NORMA.

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.

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.

Porter, M. (1985). *Competitive Advantage*. [Ventaja Competitiva]. EUA, Nueva York, Ciudad de Nueva York: THE FREE PRESS.

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

Silge, J. y Robinson, D. (2017). *Test Mining with R. A Tidy Approach.* [Minería de Texto con R. Una aproximación ordenada]. EUA: Creative Commons. Recuperado de: http://tidytextmining.com/

Wilensky, A. (2006). *Marketing Estratégico.* (7ma Ed.). Argentina, Buenos Aires: Temas Grupo Editorial.

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

7.2. Noticias en línea, leyes y conferencias

Decreto Nº 618/97. Administración Federal de Ingresos Públicos. El Poder Ejecutivo Nacional. Ciudad Autónoma de Buenos Aires, 14 de Julio de 1997. Recuperado de: http://servicios.infoleg.gob.ar/infolegInternet/anexos/40000-44999/44432/texact.htm

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/

Hohendahl, A. (2011). Procesamiento de Lenguaje Natural Robusto. Primer Encuentro de Grupos de Investigación sobre Procesamiento del Lenguaje Homenaje a Juan Seguí. Biblioteca Nacional, Buenos Aires, Argentina. Recuperado de: https://www.researchgate.net/publication/257111425_Procesamiento_de_Lenguaje_Natural_Robusto

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

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

Hinojosa, V. (2015, 10 de junio). Tres cambios que están transformando al sector hotelero mundial. *HOSTELTUR*. Recuperado de: http://www.hosteltur.com/111461_tres-cambios-estan-transformando-sector-hotelero-mundial.html

Ley Nº 18.828. Ley Nacional de Hotelería. El Presidente de la Nación Argentina. Buenos Aires, 06 de Noviembre de 1970. Recuperado de: http://wwww.observatur.edu.ar/index2.php?option=com_content&do_pdf=1&id=54

Ley Nº 4.701. Ley de Regulación de Alojamientos Turísticos. La Legislatura de la Ciudad Autónoma de Buenos Aires. Buenos Aires, 06 de Julio de 2013. Recuperado de: http://www2.cedom.gov.ar/es/legislacion/normas/leyes/ley4631.html

Rifai, T. (2014). Online Guest Reviews and Hotel Classification Systems. An Integrated Approach. [Reseñas de los huéspedes en línea y sistemas de clasificación de hoteles. Una aproximación integrada]. UNWTO, Madrid, España. Recuperado de: http://cf.cdn.unwto.org/sites/all/files/pdf/online_guest_reviews_and_hotel_classification_sytems_an_integrated_approach.pdf

7.3. Programas y extensiones

Arnholt, A. (2012). BSDA: Basic Statistics and Data Analysis. R package version 1.01. Recuperado de: https://CRAN.R-project.org/package=BSDA

Caeiro, F. y Mateus, A. (2014). randtests: Testing randomness in R. R package version 1.0. Recuperado de: https://CRAN.R-project.org/package=randtests

Duncan Temple Lang and the CRAN Team (2016). RCurl: General Network (HTTP/FTP/...) Client Interface for R. R package version 1.95-4.8. Recuperado de: https://CRAN.R-project.org/package=RCurl

Duncan Temple Lang and the CRAN Team (2016). XML: Tools for Parsing and Generating XML Within R and S-Plus. R package version 3.98-1.5. Recuperado de: https://CRAN.R-project.org/package=XML

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

Hahsler, M., Buchta, C., Gruen, B. y Hornik, K. (2016). arules: Mining Association Rules and Frequent Itemsets. R package version 1.5-0. Recuperado de: https://CRAN.R-project.org/package=arules

Hahsler, M. y Chelluboina, S. (2016). arulesViz: Visualizing Association Rules and Frequent Itemsets. R package version 1.2-0. Recuperado de: https://CRAN.R-project.org/package=arulesViz

Harrison, J. (2017). RSelenium: R Bindings for 'Selenium WebDriver'. R package version 1.7.1. Recuperado de: https://CRAN.R-project.org/package=RSelenium

Hennig, C. (2015). fpc: Flexible Procedures for Clustering. R package version 2.1-10. Recuperado de: https://CRAN.R-project.org/package=fpc

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

Karatzoglou, A., Smola, A., Hornik, K., y Zeileis A. (2004). kernlab - An S4 Package for Kernel Methods in R. Journal of Statistical Software 11(9), 1-20. Recuperado de: http://www.jstatsoft.org/v11/i09/

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., y Hornik, K. (2016). cluster: Cluster Analysis Basics and Extensions. R package version 2.0.5. Recuperado de: https://cran.r-project.org/package=cluster

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

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/

Ren, K. (2016). rlist: A Toolbox for Non-Tabular Data Manipulation. R package version 0.4.6.1. Recuperado de: https://CRAN.R-project.org/package=rlist

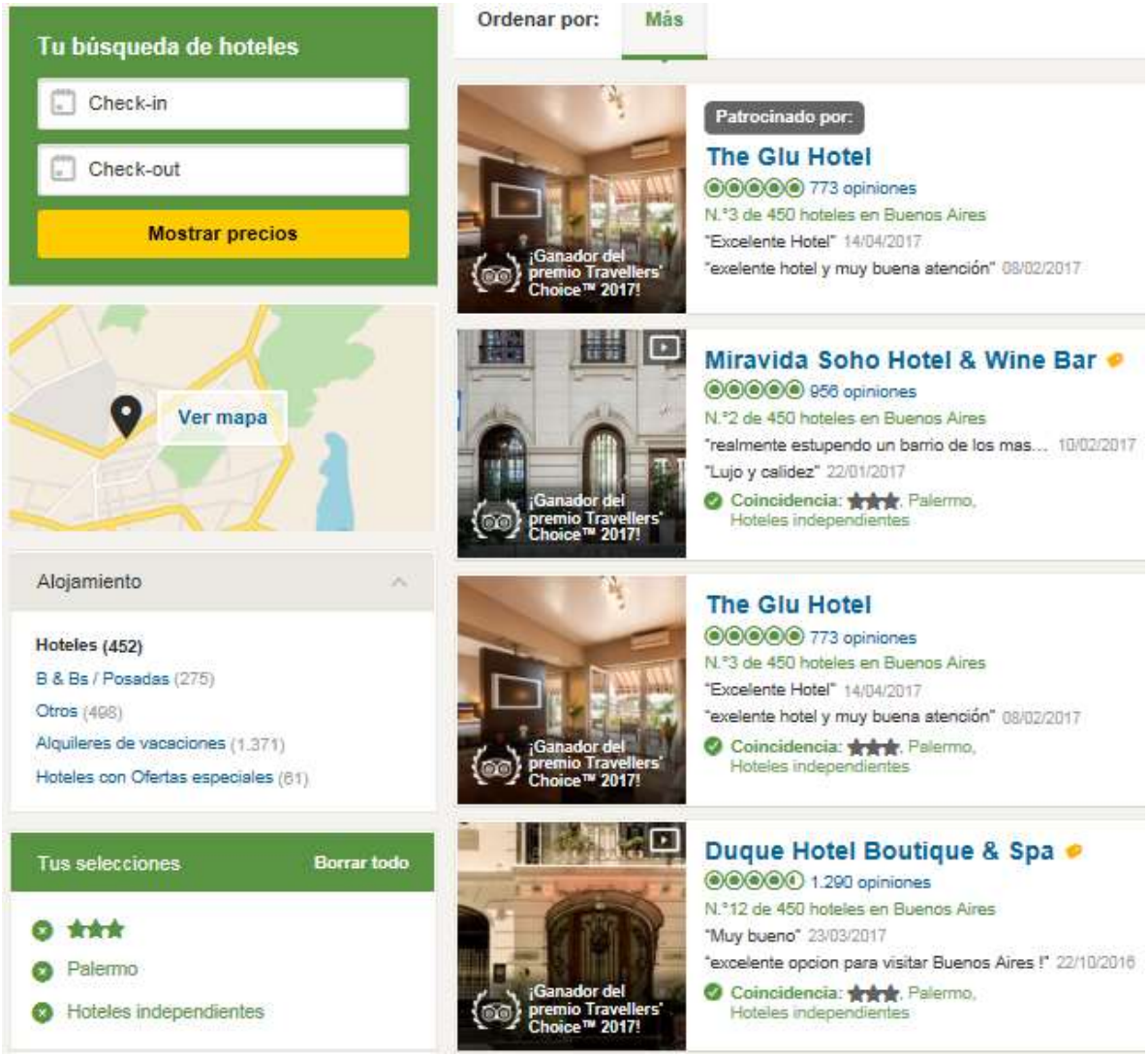
Wei, T. y Simko, V. (2016). corrplot: Visualization of a Correlation Matrix. R package version 0.77. Recuperado de: https://CRAN.R-project.org/package=corrplot

Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. Recuperado de: https://CRAN.R-project.org/package=ggplot2

Wickham, H. (2016). rvest: Easily Harvest (Scrape) Web Pages. R package version 0.3.2. Recuperado de: https://CRAN.R-project.org/package=rvest

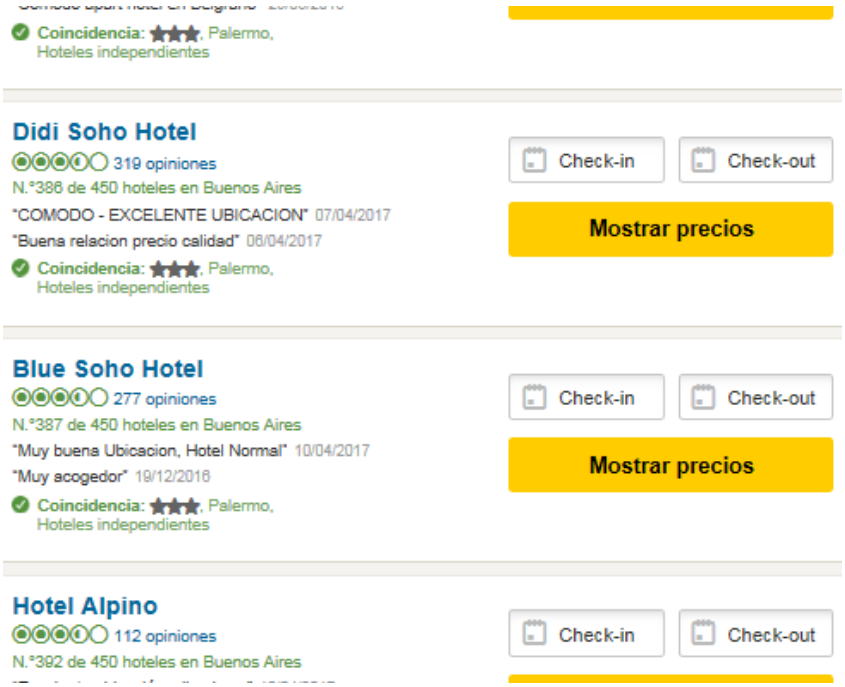
SECCIÓN DE ANEXOS

8.1. Captura de pantalla de los filtros aplicados



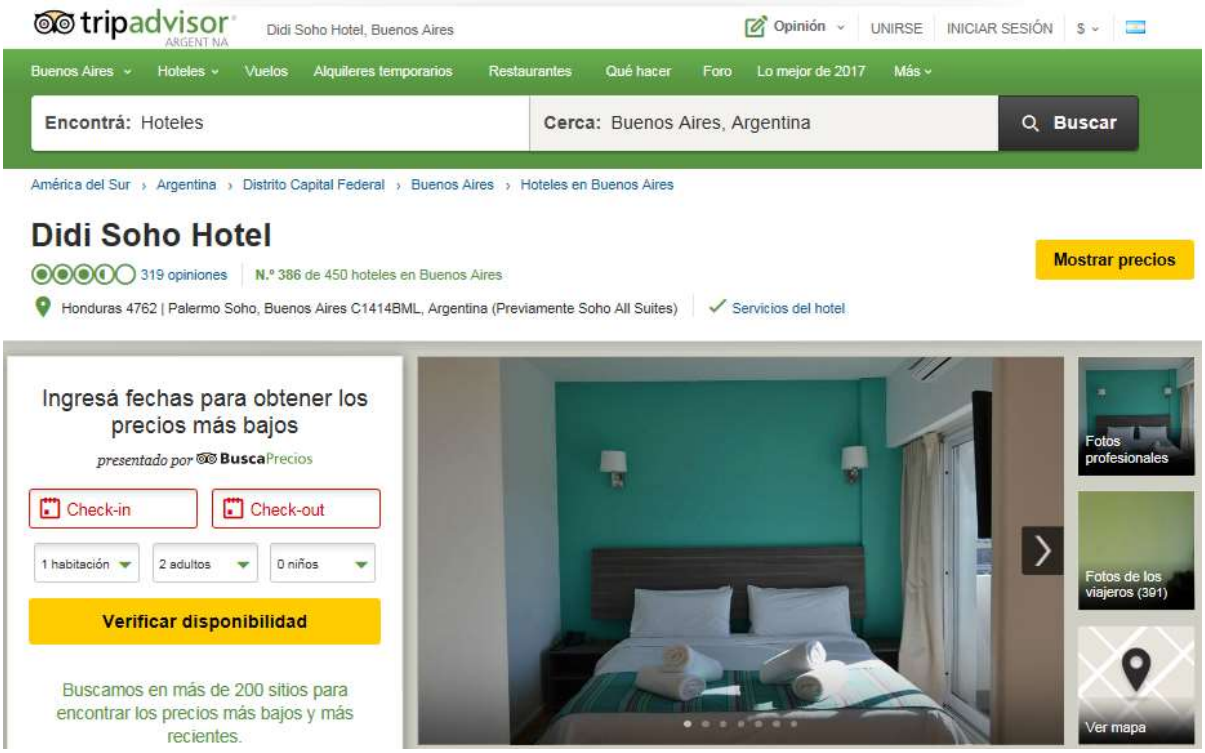
Fuente: imagen tomada de la página TripAdvisor.com el día 01 de Marzo de 2017.

8.2. Captura de pantalla de la posición de los hoteles

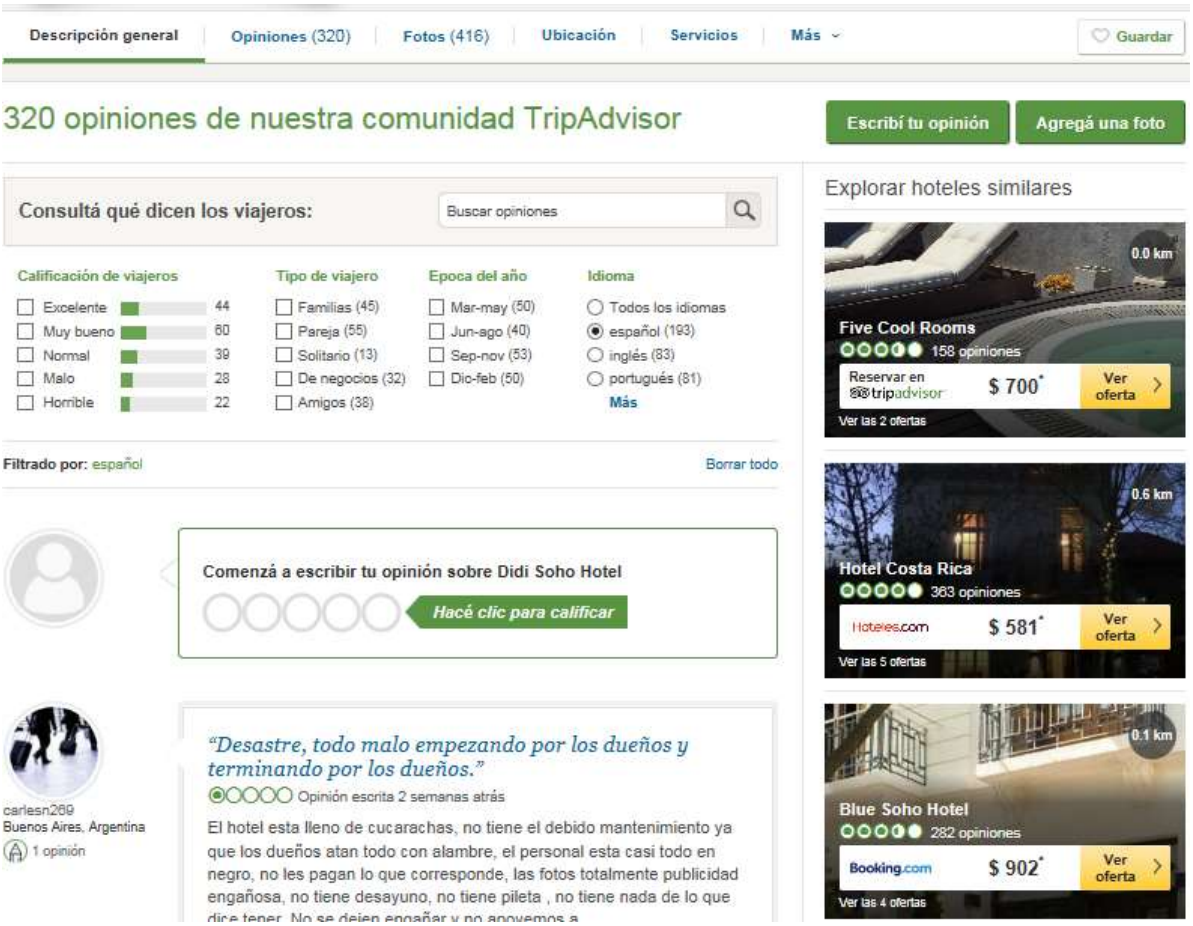


Fuente: imagen tomada de la página TripAdvisor.com el día 01 de Marzo de 2017.

8.3. Capturas de pantalla de Didi Soho Hotel



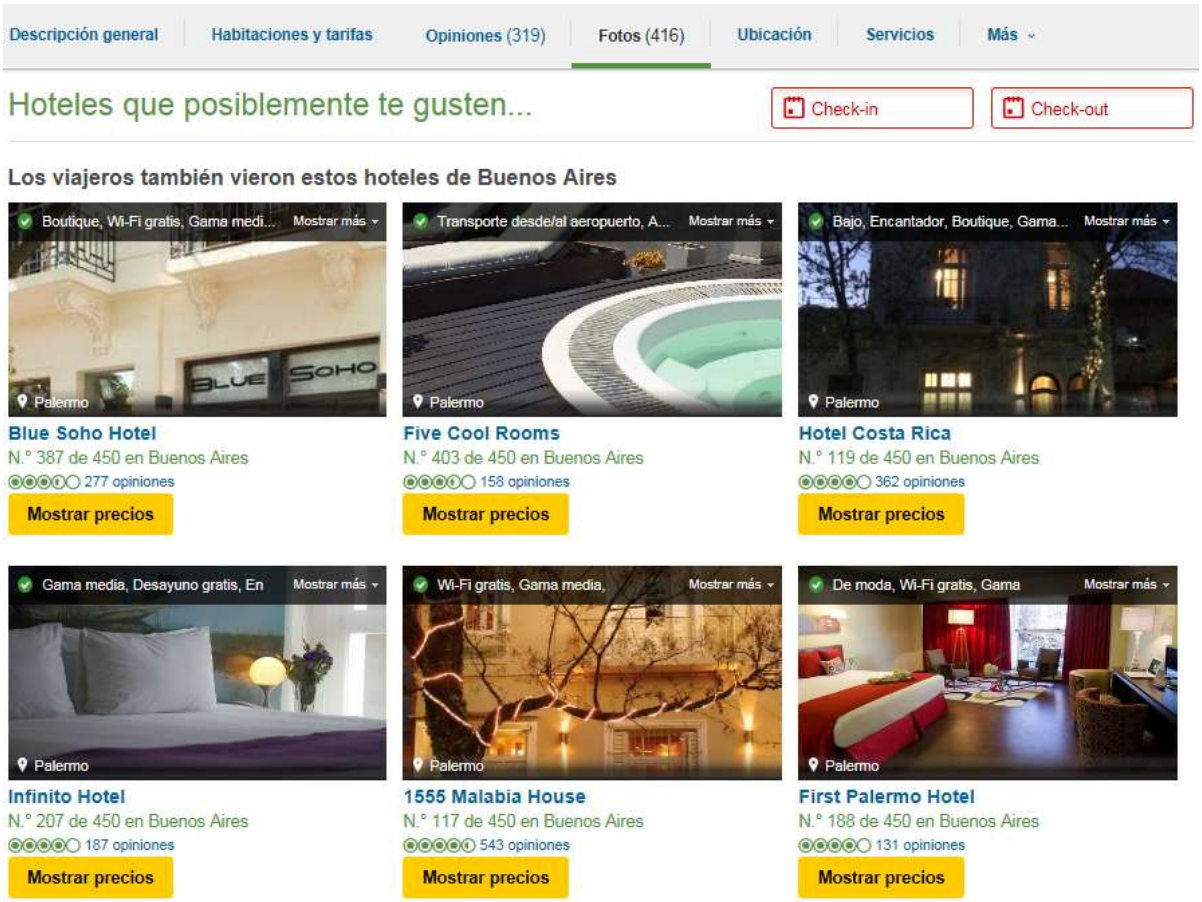
Fuente: imagen tomada de la página TripAdvisor.com el día 01 de Marzo de 2017.



Fuente: imagen tomada de la página TripAdvisor.com el día 01 de Marzo de 2017.

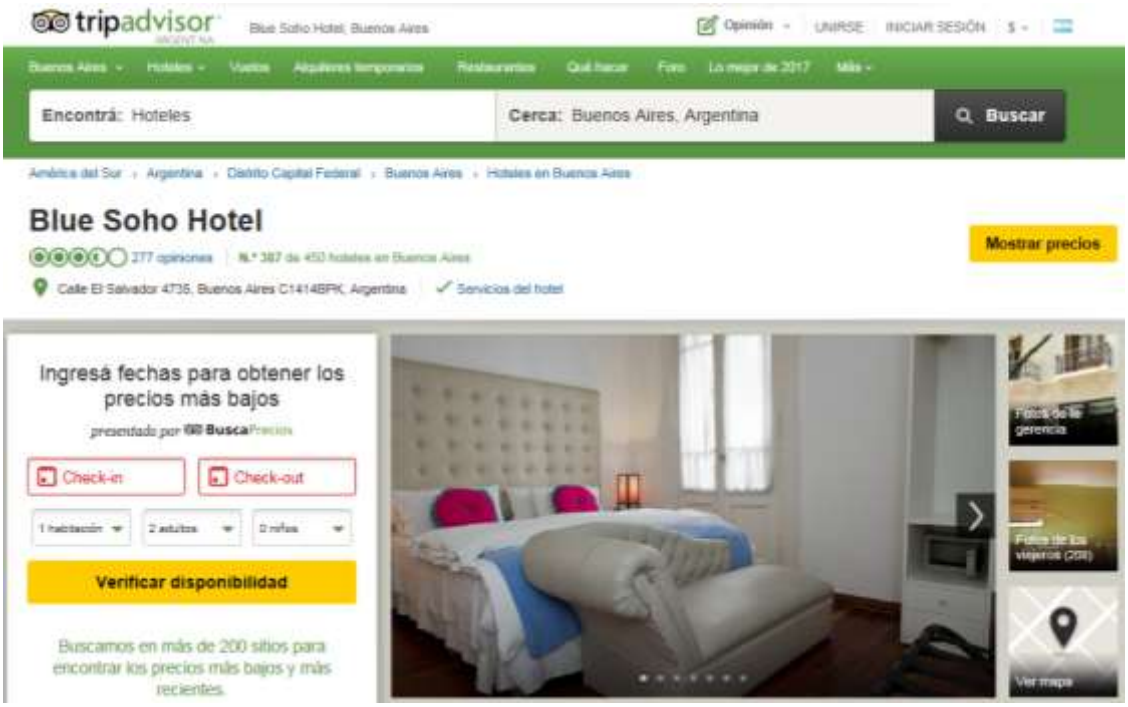
Fuente: imagen tomada de la página TripAdvisor.com el día 03 de Mayo del 2017.

Fuente: imagen tomada de la página TripAdvisor.com el día 03 de Mayo del 2017.

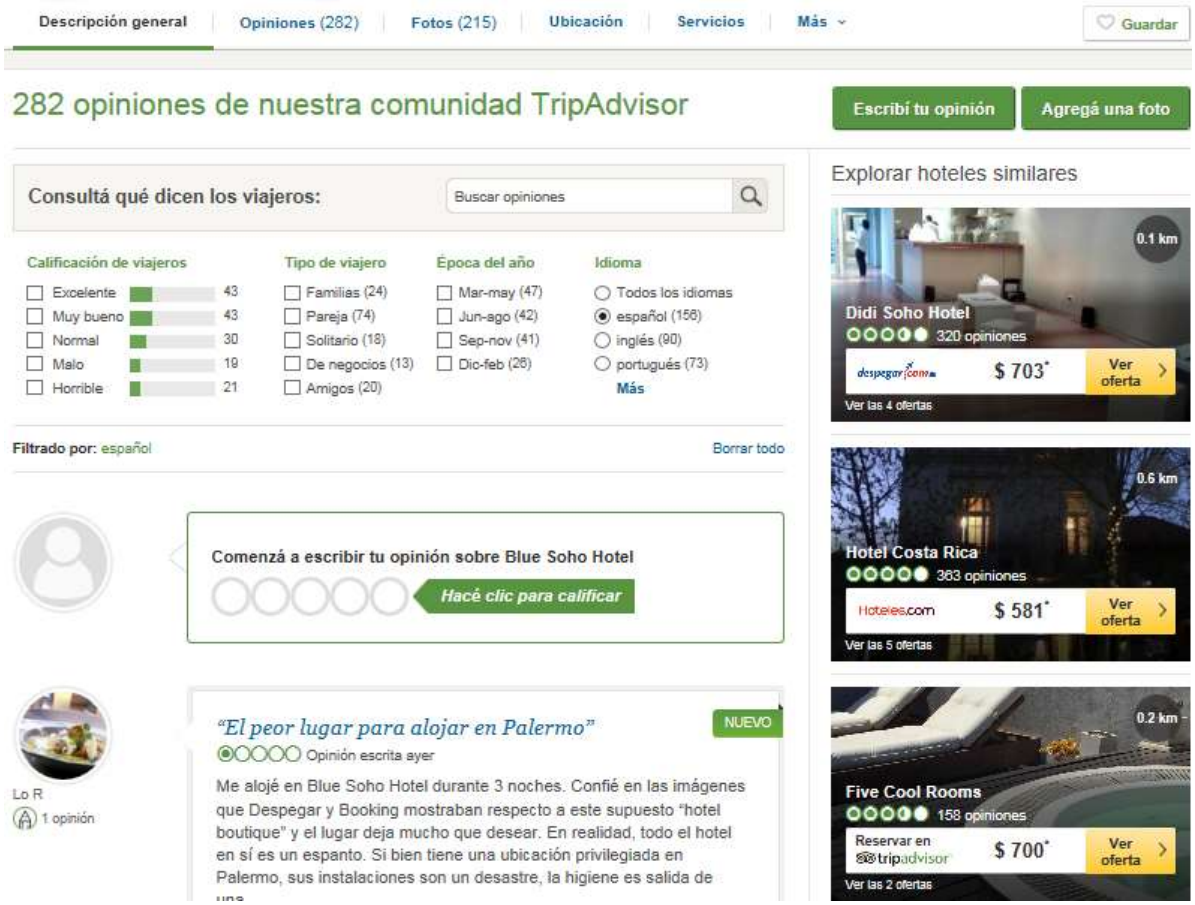


Fuente: imagen tomada de la página TripAdvisor.com el día 01 de Marzo de 2017.

8.4. Capturas de pantalla de Blue Soho Hotel



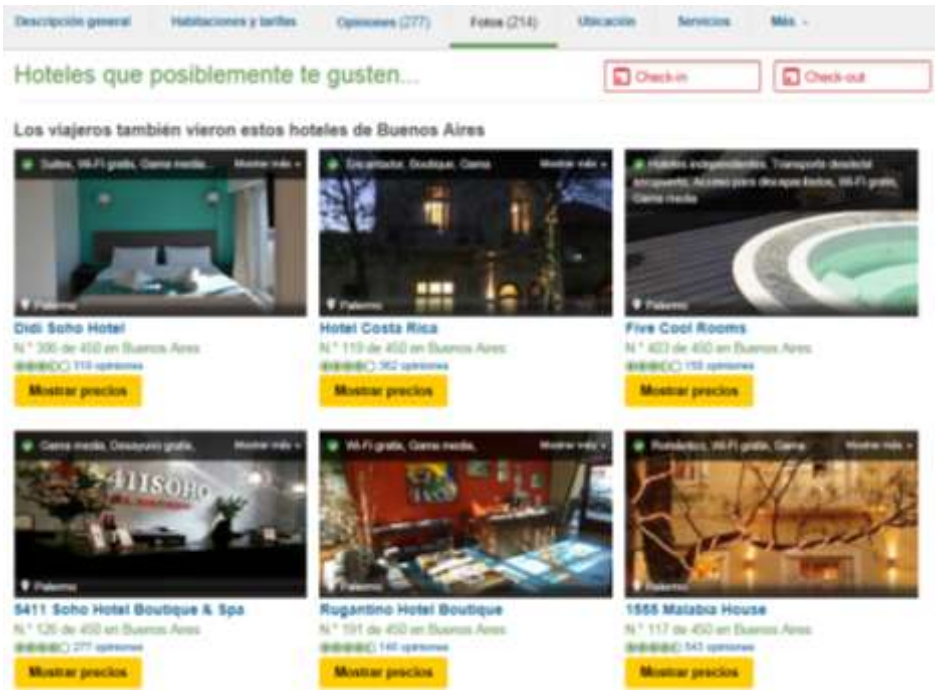
Fuente: imagen tomada de la página TripAdvisor.com el día 01 de Marzo de 2017.



Fuente: imagen tomada de la página TripAdvisor.com el día 03 de Mayo de 2017.

Fuente: imagen tomada de la página TripAdvisor.com el día 03 de Mayo del 2017.

Fuente: imagen tomada de la página TripAdvisor.com el día 03 de Mayo del 2017.



Fuente: imagen tomada de la página TripAdvisor.com el día 01 de Marzo de 2017.

8.5. Listado de frecuencias para wordcloud – Didi Soho Hotel[64]

bien	144	cerca	26	eléctrica	11	toalla	7	volvería	6
habitación	125	bar	25	pava	11	único	7	acorde	5
ubicación	112	cama	25	plaza	11	agua	6	alojamos	5
hotel	100	desayuno	25	corazón	10	atento	6	aunque	5
atención	78	noche	23	ducha	10	básico	6	calle	5
cómodo	56	sucio	20	relación	10	caro	6	cuarto	5
excelente	51	chico	19	detalle	9	check	6	didi	5
grande	46	precio	19	música	9	dejan	6	equipada	5
baño	36	boliche	18	opción	9	departamento	6	estadía	5
palermo	35	restaurant	18	ropa	9	foto	6	fiesta	5
muy	34	soho	18	pena	8	mampara	6	gente	5
recomendable	34	limpio	15	aire	8	mismo	6	ideal	5
lindo	33	local	15	ascensor	8	momento	6	imposible	5
lugar	33	lado	14	serrano	8	pared	6	luz	5
mejor	30	mantenimiento	14	viejo	8	poca	6	mañana	5
ruido	30	cuadra	13	acondicionado	7	punto	6	marca	5
mal	29	café	12	amiga	7	sabana	6	mejorar	5
amable	28	descansar	12	cocina	7	salir	6	mesa	5
limpieza	27	feo	12	falta	7	semana	6	necesario	5
zona	27	piso	12	problema	7	tipo	6	normal	5

Fuente: elaboración propia en base a resultados de R.

8.6. Listado de frecuencias para wordcloud – Blue Soho Hotel[65]

bien	139	foto	25	gente	12	pasar	8	boutique	6
hotel	130	viejo	25	mantenimiento	12	peor	8	calle	6
atención	112	recomendable	23	feo	11	perfecto	8	caro	6
habitación	104	ducha	22	precio	11	cosa	7	casona	6
ubicación	82	estafa	22	siempre	11	cuarto	7	check	6
amable	67	local	21	único	11	disfrutar	7	chic	6
muy	50	zona	20	corazón	10	gracias	7	cortina	6
desayuno	45	sucio	19	puerta	10	gusto	7	demás	6
excelente	45	baño	17	verdad	10	hora	7	dieron	6
soho	44	experiencia	17	bar	9	ideal	7	encanto	6
lindo	43	restaurant	17	barato	9	mano	7	estancia	6
cómodo	41	blue	16	cuadra	9	momento	7	estilo	6
lugar	37	descansar	15	pasillo	9	pagina	7	fin	6

[64] Sólo las 100 primeras más frecuentes.

[65] Ídem anterior.

palermo	37	detalle	15	agua	8	piso	7	inmejorable	6
noche	32	problema	15	ambiente	8	pleno	7	instalación	6
chico	31	cama	14	decoración	8	realidad	7	lado	6
ruido	29	casa	14	limpio	8	ventana	7	mismo	6
mejor	28	humedad	13	llegamos	8	además	6	opción	6
grande	26	cerca	12	llegar	8	atento	6	pena	6
mal	26	falta	12	parte	8	barrio	6	persona	6

Fuente: elaboración propia en base a resultados de R.

8.7. Listado de términos suprimidos

TÉRMINOS SUPRIMIDOS	TÉRMINOS DERIVADOS[66]
Argentina	Gentilicios, géneros y plurales
Buenos Aires	Abreviaciones
Cada	Plurales
Caso	Plurales
Cosa	Plurales
Cuenta	Plurales
Dar	Conjugaciones
Día	Número y faltas ortográficas
Haber	Conjugaciones
Hacer	Conjugaciones
Luego	(ninguno)
Poder	Conjugaciones
Ser	Conjugaciones
Solo	Género y plurales
Todo	Género y plurales
uno, dos, tres, …	Sucesivamente hasta el “diez”
Ver	Conjugaciones
Vez	Plurales

Fuente: elaboración propia en base a resultados de R.

8.8. Listado de modificaciones de cadenas de caracteres

FORMA INICIAL[67] [68]	FORMA FINAL
amabl_, cálid_, cálid_, cordial_ amabil_, amábil_, amistos_	amable

[66] Los términos derivados son aquellos que provienen de los términos suprimidos, por ejemplo, de “Argentina” se derivan “argentino”, “argentinos”, “argentina” (sin mayúscula), y “argentinas”.

[67] El carácter “_” (guión bajo) significa que puede ser precedido y/o sucedido por otro(s) caracter(es) que no sea el espacio.

[68] Las modificaciones pueden ser del tipo ortográfica, de raíz o por similitud semántica.

personal, servicio, recepción, recepcionista_, encargad_, mucama, staff, trato	atención
bano, banno, banio, ba;o, bao	baño
barat_, economic_	barato
buen_	bien
pub, disco, discoteca, boliches	boliche
cara, caras, caros	caro
cómod_, comod_, confort_, acogedor_	cómodo
detall_	detalle
dormer, descans_	descansar
_ciones, _cion, _cíon, _cín, _cin	_ción
electric_	eléctrica
engañ_, estafa_, mentir_, mintier_, miente_, decepci_	estafa
fea, feos, feas, feísim_, feisim_	feo
ampli_, espacio_	grande
alojamiento, edificio	hotel
limpia, limpias, limpios, limpit_	limpio
lind_, bonit_, hermos_	lindo
mal_, malisim_, malísim_	mal
musica	música
much_, bastante_, demasiad_, realmente, abunda_, tan	muy
madrugad_, noches	noche
recomiend_, recomend_	recomendable
ruid_, escuch_	ruido
suci_, mancha_, asco, asque_	sucio
ubicad_	ubicación
viej_, antigu_, antigü_	viejo
_zon	_zón

Fuente: elaboración propia en base a resultados de R.

8.9. Listado de términos para la identificación de tópicos

CONJUNTOS DE TÉRMINOS	TÓPICOS
“limpio”, “sucio”, “asqueroso”, “asco”, “mancha”, “higiene”, “humedad”, “mantenimiento”, “roto”, “arreglo”, “deterioro”, “decoración”.	Limpieza
“precio”, “costo”, “dinero”, “pago”, “caro”, “barato”, “económico”, “tarifa”.	Relación precio-calidad
“dormir”, “descansar”, “sueño”, “ruido”, “madrugada”, “escucha”, “sonido”, “música”, “insonorización”	Calidad de descanso
“servicio”, “personal”, “atención”, “amable”, “cordial”, “atento”, “pendiente”, “recepción”, “predispuesto”.	Calidad de servicio

“habitación”, “cómodo”, “grande”, “amplio”, “cama”, ”confort”, “espacio”.	Habitaciones
“ubicación”, “zona”, “cuadra”, “barrio”, “localizado”, “calle”, “cerca”.	Ubicación
“wifi”, “wi-fi”, “online”, “on-line”, “internet”, “conexión”, “fotos”, “página”, “web”.	Tecnología
“heladera”, “plato”, “TV”, “control remoto”, “teléfono”, “microonda”, “pava”, “instalaciones”, “equipamiento”, “vajilla”, “aire acondicionado”, “taza”, “plato”, “utensilios”, “cocina”.	Equipamiento

Fuente: elaboración propia en base a resultados de R.

8.10. Listado de términos de coocurrencia con “ubicación”

excelente	47%	serrano	14%	grande	8%	bar	6%
bien	32%	ropa	12%	soho	8%	plaza	6%
mejor	17%	cerca	11%	limpio	7%	precio	5%
palermo	17%	corazón	11%	lugar	7%	relación	5%
local	15%	restaurant	10%	muy	7%		
hotel	14%	cuadra	8%	unico	7%		

Fuente: elaboración propia en base a resultados de R.

8.11. Listado de términos de coocurrencia con “habitación”

grande	43%	baño	17%	atención	12%	limpio	8%
cómodo	27%	muy	17%	bien	11%	mal	8%
hotel	21%	desayuno	16%	limpieza	11%	sucio	6%
cama	19%	lindo	16%	café	10%		
amable	18%	ruido	16%	descansar	10%		
chico	18%	boliche	13%	noche	9%		

Fuente: elaboración propia en base a resultados de R.

8.12. Listado de términos de coocurrencia con “atención”

amable	31%	habitación	12%	excelente	9%	grande	7%
parte	26%	limpieza	12%	limpio	9%	persona	7%
atento	23%	didi	11%	viejo	8%	puerta	7%
bien	21%	gente	11%	baño	7%	mismo	6%
check	16%	hotel	11%	chico	7%	muy	6%
mal	13%	mejorar	11%	detalle	7%		
problema	13%	departamento	10%	equipada	7%		

Fuente: elaboración propia en base a resultados de R.

8.13. Listado de términos de coocurrencia con “limpieza”

mejorar	41%	amable	13%	excelente	8%	básico	6%
puerta	24%	atención	12%	feo	8%	dejan	6%
toalla	24%	falta	12%	cama	7%	semana	6%
punto	21%	habitación	11%	ideal	7%	ascensor	5%
detalle	17%	bien	10%	mantenimiento	7%	baño	5%
servicio	15%	cómodo	9%	muy	7%	unico	5%
pared	14%	hotel	9%	normal	7%		

Fuente: elaboración propia en base a resultados de R.

8.14. Listado de términos de coocurrencia con "mantenimiento"

falta	29%	baño	14%	muy	9%	limpieza	7%
detalle	25%	mejorar	11%	único	9%	feo	6%
mal	17%	puerta	11%	amable	7%	piso	6%
normal	11%	pared	10%	ducha	7%		

Fuente: elaboración propia en base a resultados de R.

8.15. Listado de términos de coocurrencia con "precio"

Acorde	30%	cualquier	10%	amable	9%	problema	7%
Pareció	22%	experiencia	10%	servicio	9%	desayuno	6%
Relación	20%	instalación	10%	momento	8%	opción	6%
Zona	18%	mal	10%	salir	8%	mejor	5%
Bien	14%	pasar	10%	volvería	8%	ubicación	5%

Fuente: elaboración propia en base a resultados de R.

8.16. Listado de términos de coocurrencia con "descansar"

música	28%	boliche	10%	eléctrica	7%
noche	20%	habitación	10%	pava	7%
ruido	17%	bar	9%	lado	6%

Fuente: elaboración propia en base a resultados de R.

8.17. Listado de términos de coocurrencia con "baño"

mojado	46%	departamento	17%	mantenimiento	14%	sabana	11%
ducha	38%	foto	17%	nunca	14%	falta	10%
detalle	35%	habitación	17%	pelo	14%	agua	9%
mampara	32%	pared	17%	volvería	14%	ascensor	9%
sucio	25%	grande	16%	amable	13%	check	9%
chico	21%	hicimos	14%	luz	12%	toalla	9%
piso	20%	lindo	14%	parte	12%		

puerta	19%	llegamos	14%	feo	11%

Fuente: elaboración propia en base a resultados de R.

8.18. Asignación de términos "positivo", "negativo" y "potenciado"

TÉRMINOS QUE PUEDEN DENOTAR SENTIMIENTO	TÉRMINO EXTRA A ASIGNAR	OBSERVACIONES
"bien", "bárbaro", "excelente", "perfecto", "impecable", "amable", "cálido", "cordial", "agradable", "amistoso", "limpio", "lindo", "bonito", "recomendable", "cómodo", "confortable", "acogedor", "amplio", "grande", "barato", "económico", "gracias".	"positivo"	En ambos casos se consideraron los mismos términos derivados, sinónimos y las faltas ortográficas de pasos anteriores para asignar el término extra
"no", "mal", "feo", "desastre", "horrible", "terrible", "deteriorado", "ruido", "roto", "agujero", "sucio", "viejo", "antiguo", "estafa", "mojado", "caro", "decepción", "mentira", "engaño", "fraude".	"negativo"	
"muy", "mucho", "bastante", "demasiado", "realmente", "abundantemente", "tan", "tanto", "extremadamente".	"potenciado"	

Fuente: tabla creada por el investigador.

8.19. Índices de coocurrencia para el análisis del sentimiento

	ascensor	atención	baño	bar	boliche
negativo	4.90%	14.60%	33.32%	11.01%	9.91%
positivo	-2.72%	22.65%	15.00%	10.26%	-0.54%

	café	cama	cocina	desayuno	descansar
negativo	5.24%	7.71%	-3.29%	13.29%	18.79%
positivo	-0.48%	30.70%	7.07%	2.72%	4.15%

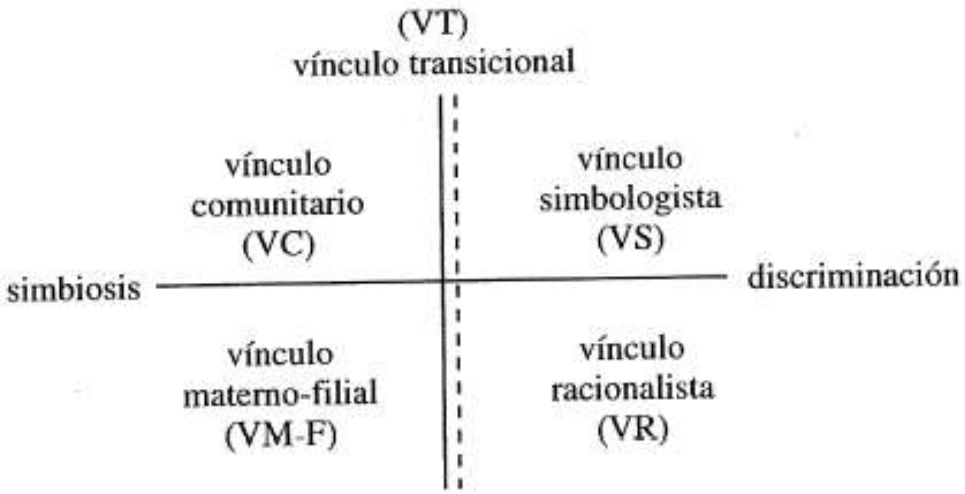
	detalle	ducha	habitación	hotel	limpieza
negativo	17.06%	15.58%	27.24%	23.67%	14.10%
positivo	5.89%	18.01%	39.94%	18.80%	8.67%

	mantenimiento	música	noche	piso	potenciado
negativo	8.48%	31.77%	15.11%	14.53%	12.52%
positivo	0.79%	-2.17%	-0.02%	-1.85%	34.11%

	precio	problema	relación	ruido	ubicación
negativo	5.06%	29.16%	1.60%	37.91%	-8.96%
positivo	6.85%	2.16%	8.16%	-1.48%	30.45%

Fuente: elaboración propia en base a resultados de R.[69]

8.20. El modelo de segmentación vincular



Fuente: gráfico tomado del texto de Wilensky (2006, p. 85)

8.21. Las reglas de asociación generadas

Antecedente		Consecuente	Soporte	Confianza	Lift
{objetividad70=1}	=>	{objetividad60=1}	11.88%	100%	4.21
{objetividad70=1}	=>	{objetividad50=1}	11.88%	100%	2.06
{objetividad70=1}	=>	{objetividad40=1}	11.88%	100%	1.77
{objetividad70=1}	=>	{objetividad30=1}	11.88%	100%	1.40
{objetividad70=1}	=>	{objetividad20=1}	11.88%	100%	1.29
{objetividad70=1}	=>	{aspecto_habitacion=1}	11.88%	100%	1.12
{subjetividad=0}	=>	{objetividad20=1}	12.87%	100%	1.29
{subjetividad=0}	=>	{aspecto_habitacion=1}	12.87%	100%	1.12
{motivo=negocios}	=>	{aspecto_tecnologia=0}	15.84%	100%	1.19
{sentimiento=negativo}	=>	{calificacion=malo}	21.78%	100%	4.04
{motivo=pareja}	=>	{aspecto_habitacion=1}	21.78%	100%	1.12
{objetividad20=0}	=>	{objetividad30=0}	21.78%	100%	3.61
{objetividad20=0}	=>	{objetividad40=0}	21.78%	100%	2.35
{objetividad20=0}	=>	{objetividad50=0}	21.78%	100%	1.98
{objetividad20=0}	=>	{objetividad60=0}	21.78%	100%	1.33
{objetividad20=0}	=>	{subjetividad=1}	21.78%	100%	1.16
{objetividad20=0}	=>	{objetividad70=0}	21.78%	100%	1.15
{objetividad60=1}	=>	{objetividad50=1}	23.76%	100%	2.06

[69] Con respecto a los porcentajes, estos fueron redondeados a dos dígitos mediante Excel.

{objetividad60=1}	=>	{objetividad40=1}	23.76%	100%	1.77
{objetividad60=1}	=>	{objetividad30=1}	23.76%	100%	1.40
{objetividad60=1}	=>	{objetividad20=1}	23.76%	100%	1.29
{objetividad30=0}	=>	{objetividad40=0}	27.72%	100%	2.35
{objetividad30=0}	=>	{objetividad50=0}	27.72%	100%	1.98
{objetividad30=0}	=>	{objetividad60=0}	27.72%	100%	1.33
{objetividad30=0}	=>	{objetividad70=0}	27.72%	100%	1.15
{aspecto_equipamiento=1}	=>	{aspecto_habitacion=1}	28.71%	100%	1.12
{objetividad40=0}	=>	{objetividad50=0}	42.57%	100%	1.98
{objetividad40=0}	=>	{objetividad60=0}	42.57%	100%	1.33
{objetividad40=0}	=>	{objetividad70=0}	42.57%	100%	1.15
{objetividad50=1}	=>	{objetividad40=1}	48.51%	100%	1.77
{objetividad50=1}	=>	{objetividad30=1}	48.51%	100%	1.40
{objetividad50=1}	=>	{objetividad20=1}	48.51%	100%	1.29
{objetividad50=0}	=>	{objetividad60=0}	50.50%	100%	1.33
{objetividad50=0}	=>	{objetividad70=0}	50.50%	100%	1.15
{objetividad40=1}	=>	{objetividad30=1}	56.44%	100%	1.40
{objetividad40=1}	=>	{objetividad20=1}	56.44%	100%	1.29
{objetividad30=1}	=>	{objetividad20=1}	71.29%	100%	1.29
{calificacion=bueno}	=>	{sentimiento=positivo}	74.26%	100%	1.29
{objetividad60=0}	=>	{objetividad70=0}	75.25%	100%	1.15
{objetividad50=0}	=>	{subjetividad=1}	49.50%	98%	1.14
{objetividad40=0}	=>	{subjetividad=1}	41.58%	98%	1.13
{subjetividad=1}	=>	{objetividad70=0}	83.17%	97%	1.11
{objetividad30=0}	=>	{subjetividad=1}	26.73%	96%	1.12
{mes=jun-ago}	=>	{aspecto_habitacion=1}	24.75%	96%	1.08
{sentimiento=positivo}	=>	{calificacion=bueno}	74.26%	96%	1.29
{objetividad60=0}	=>	{subjetividad=1}	72.28%	96%	1.12
{objetividad20=0}	=>	{aspecto_habitacion=1}	20.79%	95%	1.07
{motivo=amigos}	=>	{aspecto_habitacion=1}	20.79%	95%	1.07
{objetividad70=0}	=>	{subjetividad=1}	83.17%	95%	1.11
{anno=2014-2015}	=>	{sentimiento=positivo}	54.46%	95%	1.23

Primeras 50 reglas generadas con R según su nivel de confianza.

[i] Traducción e interpretación propia del investigador.