



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

CARRERA DE ESPECIALIZACIÓN EN MÉTODOS
CUANTITATIVOS PARA LA GESTIÓN Y ANÁLISIS
DE DATOS EN ORGANIZACIONES

TRABAJO FINAL INTEGRADOR

**Análisis de tendencias de usuarios de hoteles de
la Patagonia**

Basado en el estudio de sentimiento de texto de opiniones de turistas

AUTOR: OSVALDO ARAYA

MENTOR: MELISA ELFENBAUM

DICIEMBRE 2020



Resumen

El desarrollo de la tecnología se acrecienta a un ritmo cada vez mayor y fue modificando sustancialmente nuestra vida cotidiana. Debido a esto, la cantidad de datos que se generan en transacciones, comentarios y opiniones en redes sociales, compras, ventas, etc., crece a un ritmo exponencial, produciendo de esta manera una nueva fuente que servirá de base para obtener nuevos conocimientos. Las organizaciones, por su parte, pueden generar una ventaja competitiva obteniendo información sobre el comportamiento de sus clientes, sabiendo que es lo que piensan, cuáles son sus gustos, que prefieren comprar, dónde deciden hacerlo y demás aspectos sobre vinculados al consumo.

Por su parte, las organizaciones del rubro hotelero se mueven en escenarios muy dinámicos, cambiantes, por lo tanto, conocer los requerimientos de los usuarios se torna imprescindible para una gestión eficaz de las mismas. Específicamente, en este trabajo se analizan opiniones obtenidas de páginas webs de viajes sobre un conjunto de hoteles situados en la Patagonia Argentina y mediante técnicas de minería de texto se intenta identificar patrones o tendencias sobre la infraestructura y/o servicios que poseen dichos hoteles. El análisis de estos datos permitirá una mejor adaptabilidad al entorno y tomar decisiones de mayor calidad.

Palabras clave

Análisis de sentimiento – Minería de opinión – Opinión de usuarios – Clasificación de polaridad



Índice

1. Introducción	4
2. Gestión de datos en contextos organizacionales	7
2.1. Descripción de la organización analizada	7
2.2. Tratamiento y obtención de los datos que utiliza la organización	8
2.3. Dificultades y limitaciones vinculadas con el uso de los datos.....	13
3. Descripción metodológica	16
3.1. Recopilación de la información.....	16
3.2. Procesamiento de la información	17
3.3. Análisis de la información	18
4. Implementación de modelos de aprendizaje automático	29
4.1. Visualización.....	29
4.2. Metodologías ágiles.....	31
5. Conclusiones	33



1. Introducción

La mayoría de las personas que tomará vacaciones previamente consulta a través de la web los posibles destinos, las actividades que se pueden realizar y fundamentalmente la calidad de los hoteles mediante fotos, videos y opiniones. Principalmente en lo referido a los servicios, conocer la opinión de los consumidores se convierte en un aspecto fundamental para los administradores de los hoteles, en donde a veces las opiniones suelen ser tan dispares entre los usuarios. Por lo tanto, el desafío para estas organizaciones consiste en cómo, a partir de los datos obtenidos, se pueden traducir en información que sea útil y valiosa. Pero también surgen algunas complicaciones, la primera y más visible es que casi la mayoría de toda esta información se encuentra codificada en formato de texto, sin ninguna estructura aparente, dificultando de esta manera su tratamiento. De aquí, la necesidad de utilizar algún tipo de herramienta informática que procese este gran volumen de datos, los analice y extraiga información de forma automática.

La adaptación de los hoteles se debe a que ya no sólo alcanza con tener un sitio web o comercializar los servicios a través de agencias de viajes en la red, la aparición de la llamada Web 2.0 permite la interacción directa y la retroalimentación con el usuario, siendo su principal eje la publicación de opiniones e intercambio de informaciones (Vila, 2012). Esta nueva concepción o enfoque se sustenta principalmente en el estudio de las redes sociales de turismo como elemento clave para la comunicación y comercialización de productos turísticos (Vila, 2014). Su utilidad se resume a la extracción automática de características, lo que permite conocer la percepción que se tiene sobre aspectos concretos de un tema (Vilares, Alonso y Gómez-Rodríguez, 2013). Pero también hay que tener en cuenta el grado de veracidad de las opiniones, valoraciones vertidas, ya que la mayor parte de la información al respecto en la web puede ser "engañosa" (Perrin, 2001). En la actualidad, la mayoría de las personas consultan internet antes de decidir el destino de sus vacaciones o contratar el hotel. El uso de redes sociales y webs, son determinantes y los principales generadores de contenidos, ya en el año 2005, la cifra mundial de internautas superaba los 1.000 millones y el número sigue creciendo de manera exponencial (Sánchez Carballido, 2008). Esta "sobreabundancia" de datos también nos puede llevar a un problema al que se deben enfrentar los directivos de las organizaciones, que se produce



con la paradoja de que cada vez hay más y más información y a la vez menos tiempo para analizarla (Cano, 2007).

Ahora bien, ¿en qué consiste el análisis de sentimiento de texto? Existen diversas clasificaciones, pero todas tienden hacia una misma vertiente. Entre algunas de ellas, se puede destacar como la que se enfoca en el descubrimiento de patrones interesantes y nuevos conocimientos en un conjunto de textos (Montes, 2001). También es necesario destacar que esta metodología surge como respuesta a una nueva necesidad, tales como el auge del uso de dispositivos computacionales y de comunicación para la producción de información digital (Barrera, 2014). También puede verse como el proceso de explorar reseñas de productos de internet para determinar la opinión o el sentimiento general sobre un producto/servicio, las cuales son el contenido generado por el usuario y se transforman en un recurso para los equipos de marketing, sociólogos y psicólogos (Haddi, 2013).

Resulta interesante comprender cómo se lleva a cabo el análisis del texto, el cual consta de tres pasos: recuperación o recopilación de la información, procesamiento e integración (Kostoff & De Marco, 2001). Más específicamente, por un lado, se pone énfasis en el poder de resolución de las palabras significativas en el texto primario (Luhn, 1958) y por el otro en la organización de la información, que puede provenir del análisis de frecuencias y distribuciones de palabras (Doyle, 1961). Se conceptualiza la idea de que la literatura científica debe ser considerada como un fenómeno natural digno de exploración, correlación y síntesis (Swanson, 1988).

También suele ser llamado minería de opinión, pertenece a un área del Procesamiento de Lenguaje Natural (NLP), que también puede estar relacionado con diversas tareas como recuperación y extracción de información, traducción automática, generación de resúmenes automáticos, entre otras (Hernández & Gómez, 2013). La principal característica de la minería de texto radica en que contiene un conocimiento explícito, dependiendo de un contexto no del todo especificado, y una estructura que se puede sobreentender (Justicia de la Torre, 2017). Existen algunos problemas a la hora de analizar textos o palabras que se suman a la subjetividad y al sarcasmo, que está vinculado con pronombres y demostrativos (esto, aquello, lo último, etc.) que se refieren volviendo a otras oraciones (Johnson, Paice, Black y Neal, 1993) y con la falta de algoritmos robustos de aprendizaje automático (Bates, 1999). Es un sector que aún cuenta con mucho campo por desarrollar y se encuentra sin explotar, pero por la forma en que codifica esta



información cuenta con limitaciones para descifrar automáticamente (Hearst). En los últimos años, la minería de texto ha ganado una importancia cada vez mayor debido al gran volumen de datos de texto que

se crean continuamente, debido a esto resulta preponderante diseñar métodos y algoritmos que puedan procesar eficazmente toda esa información (Charu & ChengXiang, 2012).

La pregunta de investigación disparadora del presente trabajo es cuál es el grado de aceptación de los visitantes de los hoteles en la Patagonia y si existe cierta heterogeneidad en la calidad de los servicios prestados. La hipótesis sobre la que se parte es que existe un alto grado de aprobación (superior al 70%). El objetivo se centra en la identificación de parámetros o tendencias que permita reconocer fortalezas y debilidades sobre la infraestructura y los servicios prestados. En base a ello, será posible generar un plan de acción capaz de potenciar lo positivo y corregir o disminuir las falencias. El mismo se estructura sobre 3 (tres) aspectos fundamentales: el primero de ellos se enfoca en la gestión de los datos, en un principio se sitúa a la organización en su entorno, luego se describe todo el proceso en el cual se obtienen los datos que posteriormente serán analizados y por último se mencionan algunas limitaciones o inconvenientes que derivan del uso de esta información. En el segundo se lleva a cabo el análisis propiamente, mediante diferentes técnicas de minería de texto se obtienen métricas y comparaciones que pueden visualizarse gráficamente. Y finalmente, el último está vinculado a la implementación de los modelos de aprendizaje automático, en el cual, entre otras cosas, se describe la metodología más acorde para este tipo de proyectos.



2. Gestión de datos en contextos organizacionales

El siguiente apartado abarca una visión integral del uso de los datos y cómo los mismos se vinculan con la organización, ya sea determinando el modelo de negocio, cómo es el proceso para la obtención y cómo son utilizados luego, y los problemas que surgen de su tratamiento. Como se mencionó anteriormente, la gestión de los datos en este tipo de organizaciones es fundamental, atraviesa a gran parte de los sectores que la componen. Y por sobre todas las cosas, se vincula estrechamente con la esencia del negocio, delimitando el tipo y el alcance de las prestaciones ofrecidas.

2.1. Descripción de la organización analizada

Las organizaciones a las que se hacen referencia en el presente trabajo son las que están vinculadas al mercado turístico, más precisamente las que brindan un servicio de hotelería. En concreto, la información sobre las opiniones de los hoteles que se mencionan en este análisis refiere a las categorías de 4 (cuatro) y 5 (cinco) estrellas, las cuales cumplen con los máximos estándares de calidad. El objetivo de este tipo de empresas puede desagregarse en varios ítems los cuales pueden ser muy variados dependiendo del tamaño, el público al que apuntan y la categoría entre otras cosas, y se resumen en los siguientes: por un lado, proveer las mejores instalaciones y brindar el mejor nivel de servicios a fin de fidelizar a los visitantes manteniendo los costos lo más bajo posibles (maximizar ganancias), por otro, satisfacer en todo momento la demanda de los huéspedes, este punto enfoca tanto a la operación como a la percepción del servicio, y finalmente, generar utilidades extra ofreciendo servicios adicionales, tales como spa, restaurantes, etc.

La organización de cada hotel termina siendo muy particular al estilo y característica de cada uno de ellos, generando de esta forma múltiples variantes a la hora de analizarlas. Algunas de estas consideraciones pueden ser el segmento de clientes en los que hace foco o la filosofía de la empresa, por ejemplo, la forma de hablar del personal al relacionarse con los visitantes (vocabulario, amabilidad, etc.). A pesar de esto, existe una estructura general que puede distinguirse en cualquier organización de este tipo. Por un lado, el sector que se encarga de las habitaciones, el cual administra el servicio al



cliente, la limpieza de estas y también las reservas; el sector vinculado a la gestión de alimentos y bebidas; y después las áreas más conocidas para cualquier empresa como recursos humanos, marketing, y contabilidad.

La propuesta de valor de un hotel se basa generalmente en una investigación de mercado, que analiza las características más solicitadas o buscadas por los clientes potenciales. Es clave para los modelos de negocio de estas organizaciones definir una clasificación para las habitaciones, lo cual permite fijar diferentes rangos de precios, y, por otra parte, establecer toda la variedad de servicios complementarios que los clientes puedan llegar a consumir. También, los hoteles invierten gran parte de sus recursos a fin de lograr una fidelización de distintos operadores turísticos, sumado esto a aplicaciones y demás plataformas digitales.

Claramente los hoteles utilizan un modelo de negocio del tipo lineal. Fácilmente reconocible caracterizado básicamente por la prestación de un servicio (en forma general, hospedaje más los adicionales), tiene un lado de la transacción, es decir el visitante accede al servicio mediante una página web o aplicación que el hotel genera, y finalmente el valor se obtiene por el uso del servicio en este caso. Dicho en otras palabras, el modelo lineal crea productos y/o servicios y los vende a los consumidores. El costo marginal para el negocio lineal es decreciente hasta cierto punto, en donde comienza a aumentar exponencialmente. Aplicado a un hotel, el costo marginal de hospedar un visitante más cada vez es menor hasta el punto en donde las habitaciones estarán todas ocupadas y para seguir recibiendo turistas se deberán ampliar las instalaciones, incurriendo en una erogación de dinero muy alta.

2.2. Tratamiento y obtención de los datos que utiliza la organización

Las empresas de servicios en su mayoría se mueven en entornos complejos y volátiles. La interacción con los clientes se lleva a cabo mediante ciclos dinámicos y continuados en el tiempo, por lo que en muchos casos la información no está disponible al momento de tomar una decisión. La cantidad de información que crean los usuarios cada día servirá para una mejor toma de decisiones aprovechando los beneficios del Big Data (Sanz-Magallón Delhaize, 2018). Una de las definiciones más claras en torno a este



concepto es el que señala al Big Data como un set de datos cuyo tamaño está fuera del alcance de los softwares tradicionales para capturarlos, almacenarlos y analizarlos (Manyika et al., 2011). Para tener una idea del volumen del que estamos hablando, se estima que, a partir de este año, cada 2 (dos) días se crea tanta información como la que se creó desde el comienzo de la humanidad hasta el año 2003 (Gantz & Reinsel, 2012). De esto se desprende la importancia y lo estrechamente vinculado que se encuentra el análisis de datos con esta actividad, permite anticiparse y brinda la posibilidad de personalizar sus servicios en función de los requerimientos de los clientes (Minghetti, 2003; Luck y Stephenson, 2009). Los hoteles que se mencionan en este trabajo se ubican dentro del nivel experimentado (en base a los distintos niveles de análisis de datos), caracterizados por ser empresas de gran tamaño, en las que el objetivo es ir más allá de una reducción de costos, sino que se busca conseguir más utilidades.

El ciclo de vida de un cliente muestra las diferentes instancias por las que atraviesa un cliente cuando es atendido en un hotel (Durán, Natarajan & Giraldo, 2011). En cada una de ellas, tiene varias interacciones con el hotel y los empleados de este, y es la principal manera de vincularse e influir sobre él. La primera etapa es la de búsqueda, en donde el futuro cliente consulta información de todo tipo sobre el hotel, como disponibilidad, precios y demás servicios complementarios. Esta consulta se puede realizar a través de la web del hotel, agencias de viajes, etc., por lo que entender las necesidades de los clientes es vital a la hora de generar la propuesta de valor. La etapa posterior es la que se produce con la llegada al hotel y el registro en el mismo, en esta parte se ofrecen otros servicios complementarios a los que el usuario podrá acceder. Durante la estadía, el uso de los bienes / servicios se transforma en una fuente de información sobre los gustos y las preferencias del consumidor, pudiendo ser utilizada más adelante para mejorar los productos ofrecidos. Y finalmente, el registro de salida es la última etapa de interacción con el cliente.

Una consideración muy importante es que el hotel tiene al menos 3 (tres) instancias en las que puede recoger información sobre sus visitantes. La primera es durante su estadía actual, a medida que va avanzando en el ciclo de vida, la segunda, surge de datos del mismo cliente cuando se alojó en el hotel en una estadía anterior. Y la tercera, que se crea en base a las estadías de otros clientes que también puede ser considerada. Para este último caso por ejemplo pueden segmentarse clientes en base a



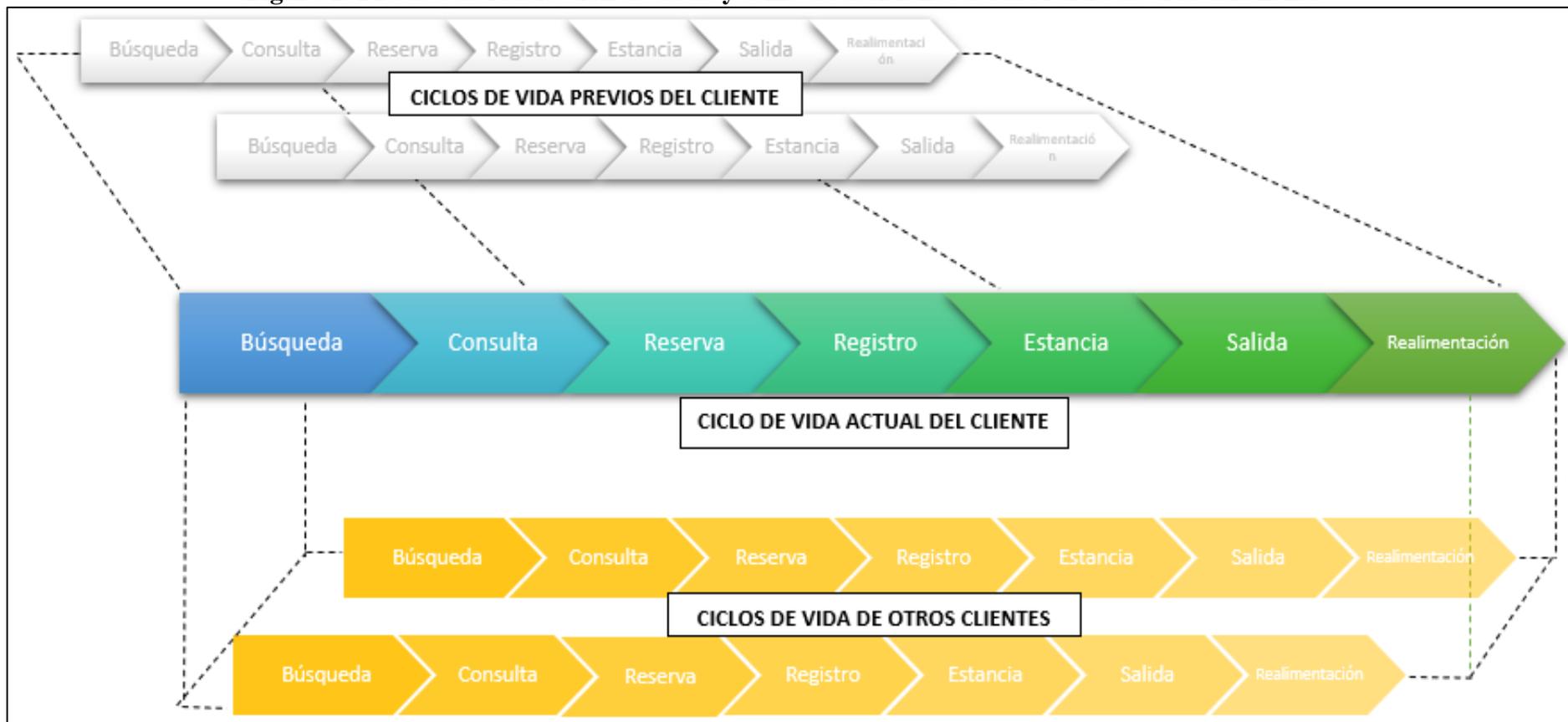
Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Estudios de Posgrado



género, edad, etc. Lo fundamental en estas organizaciones es que el proceso de recolección de datos esté sincronizado con el proceso de toma de decisiones. Si esto se produce, las decisiones que se tomen serán consecuentes con las preferencias de los visitantes, y, en consecuencia, mejores.



Figura 1: Proceso de recolección de datos y toma de decisiones desde el ciclo de vida del cliente



Fuente: Elaboración propia en base a Sincronización de los ciclos de información y de gestión de infraestructuras de servicios hoteleros



En referencia a las anteriores fuentes de datos disponibles, se construyen diferentes 4 (cuatro) tipos de decisión, vinculados estrictamente al sector hotelero (Durán, Natarajan & Giraldo, 2011). La categoría I se fundamenta en los datos obtenidos en el ciclo de vida actual del visitante. Un ejemplo de esto sería preparar un desayuno acorde a los gustos del cliente consultándole previamente su preferencia al momento de llegar y hacer el registro.

La categoría II utiliza como dato las estancias anteriores del mismo cliente, como, por ejemplo, preparar la habitación acorde a sus gustos, preferencias de colchón o almohadas, etc. Esos datos obviamente no están disponibles en el ciclo de vida actual, ya que corresponde a eventos pasados. Se realiza una extrapolación directa del comportamiento del pasado.

La categoría III combina las dos fuentes anteriores de datos, el actual ciclo y los pasados, y mediante esa combinación y mediante un análisis estadístico hace una estimación de posibles gastos que un cliente puede realizar. Para esto, utiliza datos anteriores de otros clientes sobre consumos y, según los datos actuales del cliente, lo encasilla en una clasificación. También pueden utilizarse métodos de inferencia en minería de datos, clustering o segmentación, etc.

La categoría IV tiene que ver con los datos referidos a visitas anteriores y los relaciona con información de otros clientes (similares o parecidos), suele usarse para promocionar y brindar descuentos de servicios que sean de interés para los clientes actuales. Esta categoría surge como solución ya que de antemano no se sabe la probabilidad de que un cliente use o no un servicio, ese dato se conocerá una vez finalizada la estadía. Esto es algo parecido a los sistemas de recomendación.

En resumen, los datos del ciclo de vida actual de un cliente son los mejores a utilizar a la hora de la toma de decisiones. Son los más recientes y son certeros, en este caso no hay margen de error, cumpliendo con el criterio de precisión. Pero claramente, en algunas situaciones se necesita conocer con anterioridad esta información ya que de lo contrario se dispondrá de ella en un momento posterior al de la toma de decisiones. En ese caso no cumpliría el criterio de oportuna. Para corregir esta situación, se utilizan diferentes técnicas de minería de datos y estadística en la que se extrapolan datos pasados o de otros clientes similares con los actuales (Law, Henry y Carey, 2007).



La mala calidad de la información puede traer consecuencias no muy agradables para una organización. Principalmente generará errores en la toma de decisiones, provocando informes no confiables y análisis defectuosos. Por otra parte, aumenta los costos de gestión y genera una ineficiencia operacional, y por último y quizás más importante, dificulta ofrecer un servicio de acuerdo con las necesidades de cada cliente. Las principales causas se deben a la entrada de información de forma manual (errores de tipeo, equivocaciones, duplicación de datos), a la incorporación automática de datos externos sin ser analizados, migraciones de los sistemas de gestión, entre otras.

2.3. Dificultades y limitaciones vinculadas con el uso de los datos

Como se mencionó previamente, en el tipo de organizaciones analizadas existe un desfase entre la información disponible y el momento de tomar las decisiones. Es decir, en muchos casos se dispone de los datos cuando dejan de ser útiles, ya sea porque el cliente se encuentra en otra etapa del ciclo de vida o bien porque el mismo ha dejado el hotel. Por lo tanto, para resolver este inconveniente, se establecen diferentes técnicas estadísticas para inferir, predecir, segmentar, y obtener un dato estimado o un porcentaje de ocurrencia de un evento, en base a datos pasados, de otros clientes, o ambos al mismo tiempo.

Para llevar a cabo todas estas tareas es necesario fijar pautas previas que brinden un entorno regulatorio para la obtención y la calidad de los datos. Es lo que se da en llamar *Data Governance* (traducido del inglés, Gobierno de Datos), el cual incluye las transacciones (eventos, pagos, ventas), el dato maestro o máster data (datos de cliente, proveedor, ubicación, etc.) y el *Reference Data* (datos atribuibles al máster data). Como complemento, la gestión de los datos maestros garantiza la uniformidad, precisión y coherencia de los datos. Esta información es la que se compartirá con toda la empresa, por lo tanto, no debe ser susceptible a errores. En resumen, los principales puntos hacia donde apunta el *Data Governance* son los siguientes: asegurar la consistencia de la información y que esta sea entendida por todos los implicados, establecer el uso de los datos como un activo de gran valor, mejorar la consistencia de los proyectos de la



organización y finalmente asegurar el cumplimiento de las regulaciones, tanto internas como externas (Ampuero, Alfaro, Raymundo y Domínguez, 2013).

Ahora bien, por otra parte, entra en juego un aspecto muy importante que muchas veces no es tenido muy en cuenta y es el de los riesgos que conlleva la gestión de los datos. Como primera medida, se debe analizar el grado de fiabilidad de los datos con que cuentan estas organizaciones. ¿Son confiables los datos que se analizan? Para cualquier tipo de cliente (actual o pasado), las experiencias, los servicios adquiridos, las preferencias, el hotel las almacena para utilizarlas en un futuro. Es el mismo cliente quien brinda sus datos, por lo tanto, más allá de algún error u omisión, se puede llegar a alcanzar un alto grado de fiabilidad. De aquí en más, entran en juego otras cuestiones como la veracidad o no de los mismos, aspecto que escapa de este trabajo.

Y por otro lado se encuentra el tema de la privacidad. Todas las personas tienen negocios, propósitos, o demás cuestiones de cada uno y también relacionados con su familia, que prefieren mantener en un aspecto reservado, que no sea público (Valesani, Mariño y Martínez, 2003). En cuanto al aspecto legal, existen diversas regulaciones en el mundo en cuanto al tratamiento de los datos, habiendo una marcada distinción en la legislación de Europa y Estados Unidos. En nuestro país, la ley 25326 dictada en el año 2000 tiene como objetivo “...la protección integral de los datos personales presentes en bases de datos, archivos...”. Pero también en el artículo 43 de la Constitución Nacional establece que “...toda persona podrá tomar conocimiento de los datos a ella referidos y a su finalidad”, aquí surgió la figura de habeas data. También la Capital Federal impulsó la ley 1845 relacionada con esta problemática.

Claramente, cuando un viajero completa el registro al llegar a un hotel, no sabe que esos datos podrán ser utilizados para otros fines para los cuales fueron solicitados, conocido como concepto de reasignación, según modelo de las 4R¹ (Steinmann, Matei y Collmann, 2016). También existe la posibilidad de que se extraigan datos de huéspedes de otros hoteles y se los analice para compararlos con los de los propios clientes (reutilización). Por otra parte, pueden extrapolarse y generar perfiles de clientes, obteniendo la información de diversas fuentes (recombinado). Como vemos, más allá de existir leyes y regulaciones, resulta difícil poner un límite al tratamiento de datos por parte

¹ Metodología que se utiliza para analizar la naturaleza ética que deriva del uso del Big Data



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



de las organizaciones. Los visitantes de hoteles en este caso se encuentran en una posición “muy vulnerable” en cuanto al manejo de sus datos personales. A pesar de dar consentimiento en algunas ocasiones, los hoteles disponen de ellos y no se sabe con certeza para qué, y cómo serán utilizados. Se genera aquí una paradoja, ya que el Big Data contempla y resalta a la privacidad como un valor que en muchas ocasiones no se estaría respetando.



3. Descripción metodológica

El siguiente apartado presenta el análisis propiamente dicho de las opiniones y comentarios obtenidos. Se indica en base a qué información fue generada la base de datos, las técnicas aplicadas y el software utilizado para llevarlo a cabo. También se realiza una breve descripción de los resultados obtenidos mediante distintos gráficos.

3.1. Recopilación de la información

El principal inconveniente para llevar a cabo este trabajo fue la falta de información referida a opiniones de hoteles agrupada en una base de datos. Esto se entiende en cierta manera por lo específico del tema, no se habla de hoteles en general, sólo de aquellos que están ubicados en la Patagonia Argentina. Debido a esto, surgió la necesidad de localizar, procesar y unificar la mayor cantidad de datos posibles con el objetivo que los resultados y conclusiones obtenidas reflejen la realidad de la mejor forma posible, o sea sean lo más confiable posibles.

Finalmente, fue creada una base de datos: la web elegida fue la de TripAdvisor y la técnica utilizada fue web scraping. Fue realizada mediante el complemento del navegador de Google Chrome llamado Web Scraper². Los datos obtenidos fueron volcados a una planilla de Excel en donde se hizo una operación de transformación, limpieza y se ultimaron detalles para dejar la base lista. En resumen, se obtuvieron un total de 11613 (once mil seiscientos trece) observaciones, para las que se extrajeron número de ID (identificación de la fila u observación), nombre del hotel, ubicación, opinión, ciudad de origen del turista y país de origen del turista.

La columna nombre del hotel toma 24 (veinticuatro) valores distintos, los cuales son Design Suites Calafate, Hotel Albatros, Hotel Cristal, Hotel Tirol Bariloche, Península Petit Hotel, Design Suites Bariloche, Hotel Tres Reyes, Hostería Encanto del Rio, Arakur Ushuaia, Marinas Alto Manzano, Hotel Panamericano Bariloche, NH Bariloche Edelweiss, Rochester Hotel Bariloche, Nido del Condor Hotel Spa, Rotui Apart Hotel, Pailahue Lodge Cabanas, Huinid Hotel Pioneros, Huinid Bustillo Hotel Spa,

² Herramienta gratuita de extracción de datos web, es una extensión del navegador Google Chrome.



Cacique Inacayal Lake Spa Hotel, Hotel Mirador del Lago, Águila Mora Suites Spa, Lagos del Sur Apart Suites, La Floridita Puerto Pirámides y Las Balsas. Por otro lado, la columna ubicación toma 6 (seis) valores los cuales son: El Calafate, Ushuaia, San Carlos de Bariloche, Villa La Angostura, San Martín de Los Andes y Puerto Pirámides. Las columnas ciudad y país del turista tienen valores muy dispersos, ya que los turistas provienen de cualquier parte del mundo.

Entendiendo esto, la columna opinión será el eje principal del análisis, que corresponde a la opinión propiamente dicha. Es un campo de texto, que puede contar con una o más oraciones dependiendo del caso. Es la expresión que denota la percepción de la persona que se hospedó en el hotel. Por último, vale aclarar que no existen datos faltantes (missings values), y en consecuencia de trabajar con texto (strings) no hay valores atípicos (outliers).

3.2. Procesamiento de la información

Para el tratamiento y análisis de los datos se utilizará el software RStudio en su versión 1.3.1073. En lo que concierne estrictamente a texto, resulta imprescindible realizar una limpieza de los datos previo análisis, ya que no toda la información es importante y revela datos de nuestro interés. Como primer paso, se deben eliminar los signos de puntuación (ya sea comas, puntos y acentos), exclamación e interrogación.

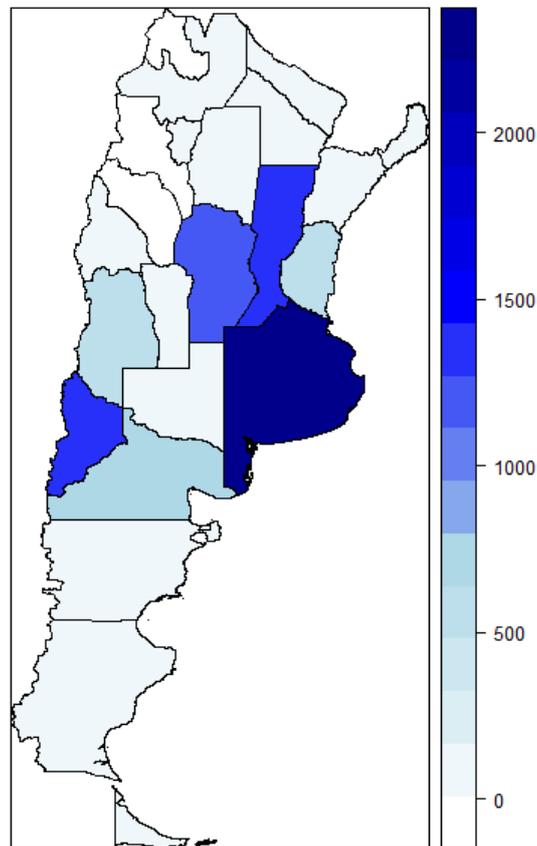
Luego se aplica una “tokenización”, que consiste en separar las oraciones en palabras, generando unidades más pequeñas para analizar. Esto se conoce también como segmentación de texto. También se eliminan aquellas palabras que no tienen un significado por sí mismas, como es el caso de preposiciones, artículos y algunos verbos no relevantes. Esta acción se lleva a cabo mediante la aplicación de “*stopwords*”, que vienen predefinidas según el idioma pero que también pueden agregarse términos que no estén incluidos. Finalmente, a modo de simplificar el análisis, se agrupan palabras similares que tienen el mismo origen, proceso llamado “*stemming*”, el cual invoca a la raíz de una palabra manteniendo el concepto y suprimiendo la cantidad de términos.



3.3. Análisis de la información

La figura 1 indica la cantidad de visitantes argentinos distribuidos por provincia, es decir, cuál es el origen de las personas de nuestro país que opinaron sobre los hoteles. Cabe mencionar que la base de datos utilizada cuenta además con opiniones de extranjeros, los cuales representan una minoría. En este sentido, la provincia de Buenos Aires, Capital Federal, Córdoba, Santa Fé y Neuquén son las que aportan un mayor número de turistas.

Figura 2: Origen de los visitantes argentinos



Según provincia de residencia

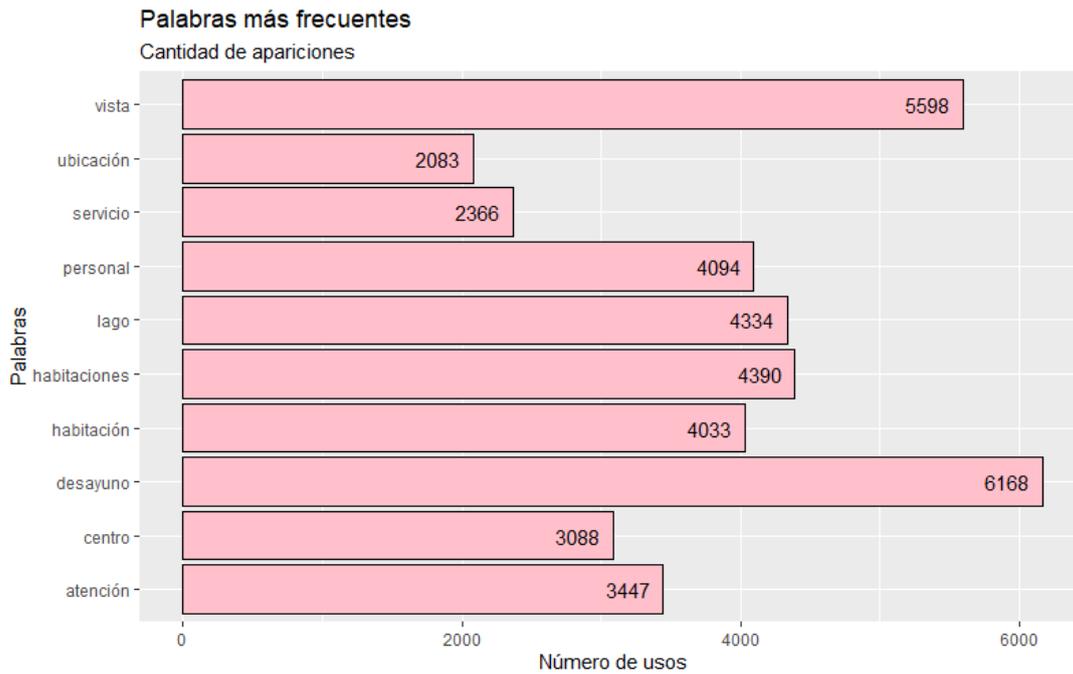
Fuente: Elaboración propia

A continuación, se realiza un conteo del total de las palabras que se encuentran en todas las opiniones. Esto dará una idea de lo que más destacan los visitantes, ya sea en cuanto a lo positivo o a lo negativo. Los resultados arrojan que la palabra más utilizada es *desayuno* con 6168 (seis mil ciento sesenta y ocho) apariciones, seguida de *vista* con



5598 (cinco mil quinientos noventa y ocho) y *habitaciones* con 4390 (cuatro mil trescientos noventa). La figura 3 indica la frecuencia de aparición, pero nada indica sobre la connotación de estas.

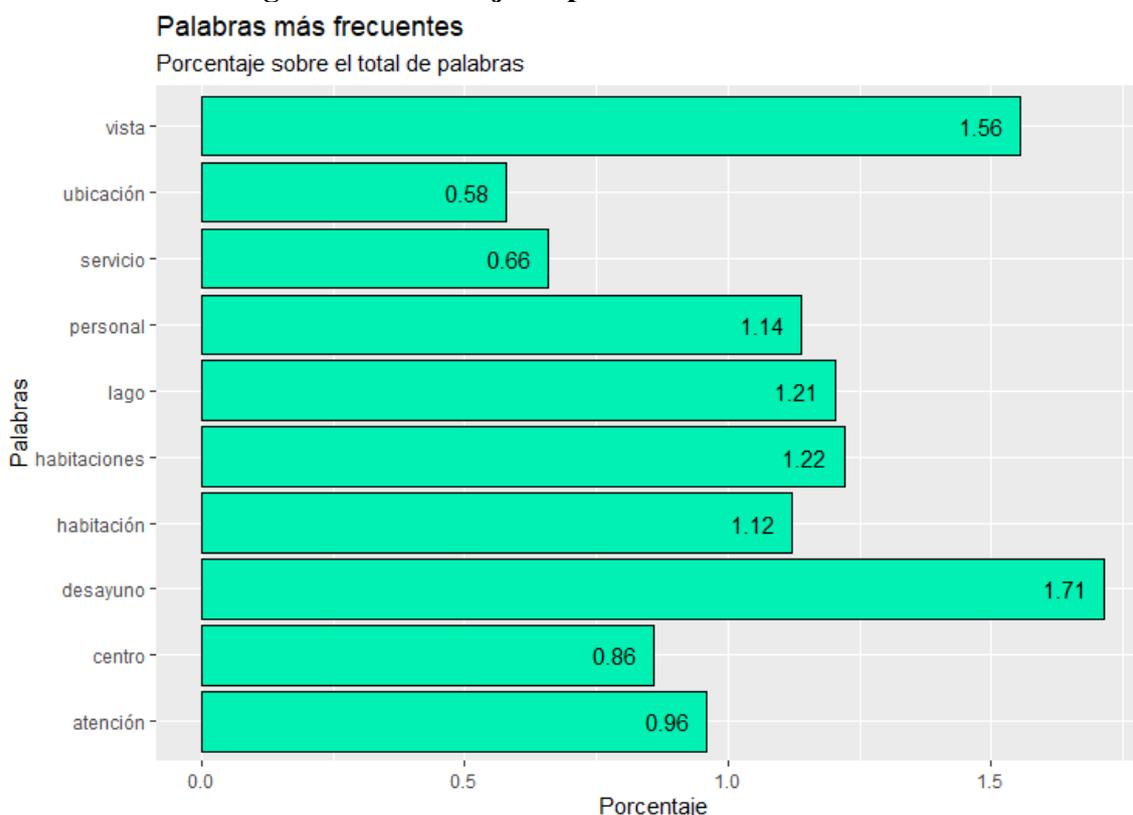
Figura 3: Palabras más frecuentes



Fuente: Elaboración propia

El gráfico anterior presenta valores absolutos, los cuales brindan una idea acerca de lo que tratan las opiniones, pero nada dice sobre el peso que representan dentro del universo de las palabras que conforman las opiniones. Para ello, la figura 4 muestra la incidencia de esas mismas palabras, las que más veces fueron utilizadas, dentro del contexto general de los datos analizados. *Desayuno*, que fue la más utilizada representa sólo el 1,71% de los términos, mientras que *vista* alcanza sólo el 1,56%. Esto demuestra la magnitud del análisis, el cual se lleva a cabo con más de 360.000 (tres cientos sesenta mil) palabras. A pesar de ser las palabras más comunes, no llegar a ser lo suficientemente representativas, esto se debe en gran parte al volumen de la base de datos analizada, que consta de más de 11.000 (once mil) opiniones.

Figura 4: Porcentaje de palabras más frecuentes



Fuente: Elaboración propia

Más allá de los gráficos, otra forma más visual para ver estos resultados, es la conocida nube de palabras (figura 5). Básicamente consiste en representar en este caso la frecuencia de los términos más utilizados asignando un tamaño y una ubicación en base a su importancia. En consecuencia, la palabra más común aparece en el centro y de un volumen mayor al resto, y así en forma decreciente con el resto de los términos. Asimismo, las que menos frecuencia tienen se ubican en la periferia de la “nube”. Se distinguen 3 (tres grupos) de palabras: las más frecuentes ya mencionadas anteriormente, otro grupo con importancia un poco menor al primero (con términos como instalaciones, comida, pileta, cabañas, spa, recepción, etc.) y finalmente todo el resto de las palabras que tienen muy pocas apariciones en las opiniones.



Es importante tener en cuenta dos consideraciones, por un lado, este diccionario es el resultante de una traducción del idioma inglés, por consiguiente, alguna palabra puede no ser del todo correcta o precisa, y, por otro, la asignación de puntajes o valores es totalmente subjetiva y queda a criterio del creador de dicho diccionario. Más allá de esto, y a pesar de las imperfecciones que se puedan producir, resulta altamente compatible para llevar a cabo este trabajo. Por ejemplo, conceptos como fraude, infierno y molesto alcanzan los máximos valores negativos, mientras que excepcional, encantado, agradable, alcanzan los máximos valores positivos.

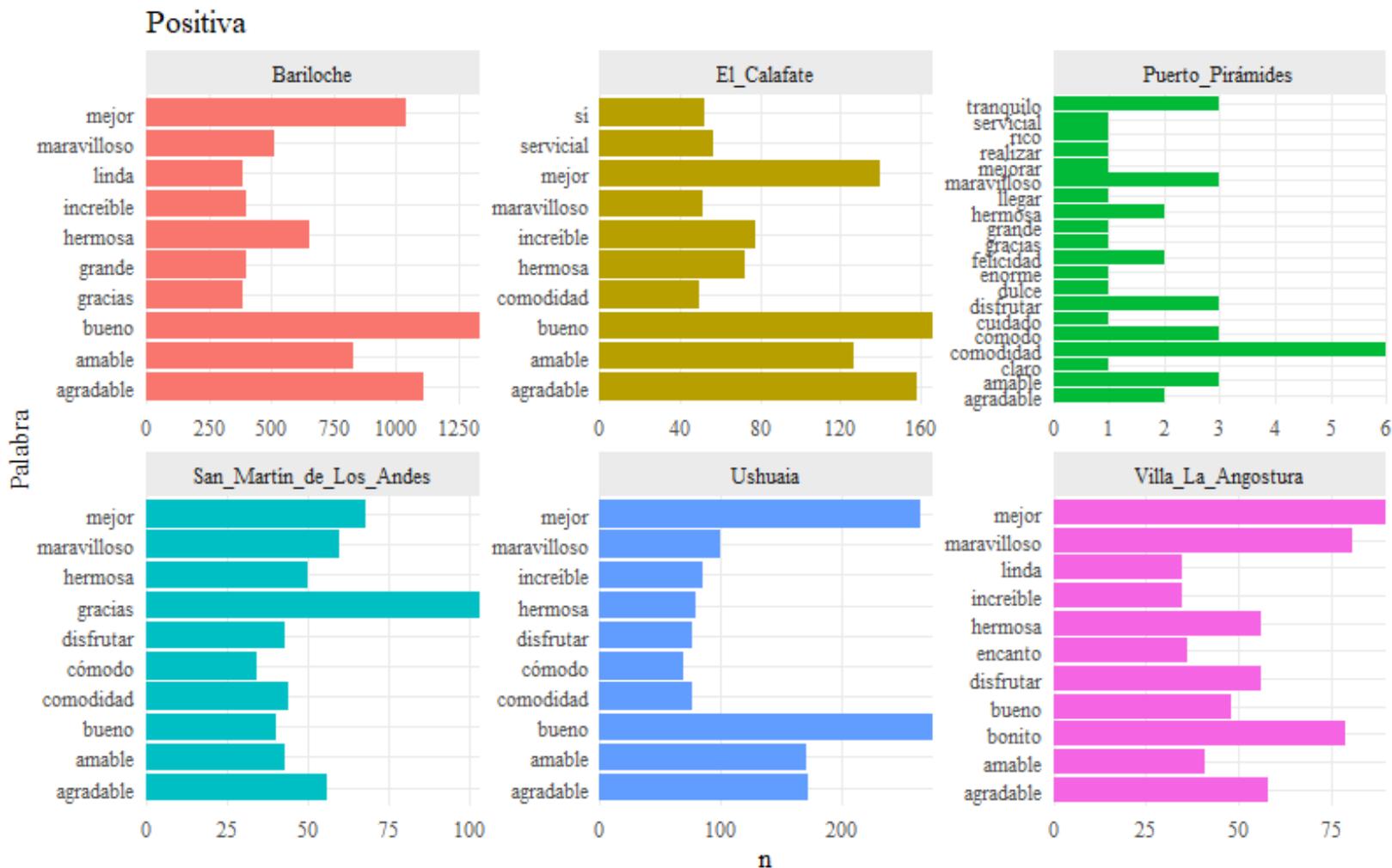
La figura 6 se centra en las opiniones positivas y abre el panorama en cuanto a la ubicación de los hoteles. Las palabras más frecuentes en este tipo de valoraciones son *bueno, agradable, comodidad, maravilloso*. Esto se reproduce y es común a todas las ubicaciones.

La figura 7 reproduce la misma información que la figura anterior, pero para las opiniones negativas. En este sentido, se omitió la palabra *no*, que es muy común para dar una connotación negativa y no aporta mucha información. Los términos más comunes y que también se repiten en todas las ubicaciones son *problema, dudas, pena*, etc.

En algunos casos, se cuenta con más información para unos hoteles que para otros, es decir, por ejemplo, la cantidad de opiniones es mucho mayor en Bariloche que en el caso de Puerto Pirámides. Igualmente, lo que demuestran estos gráficos es que, a pesar de analizar distintos establecimientos (hoteles), los datos obtenidos presentan una homogeneidad.



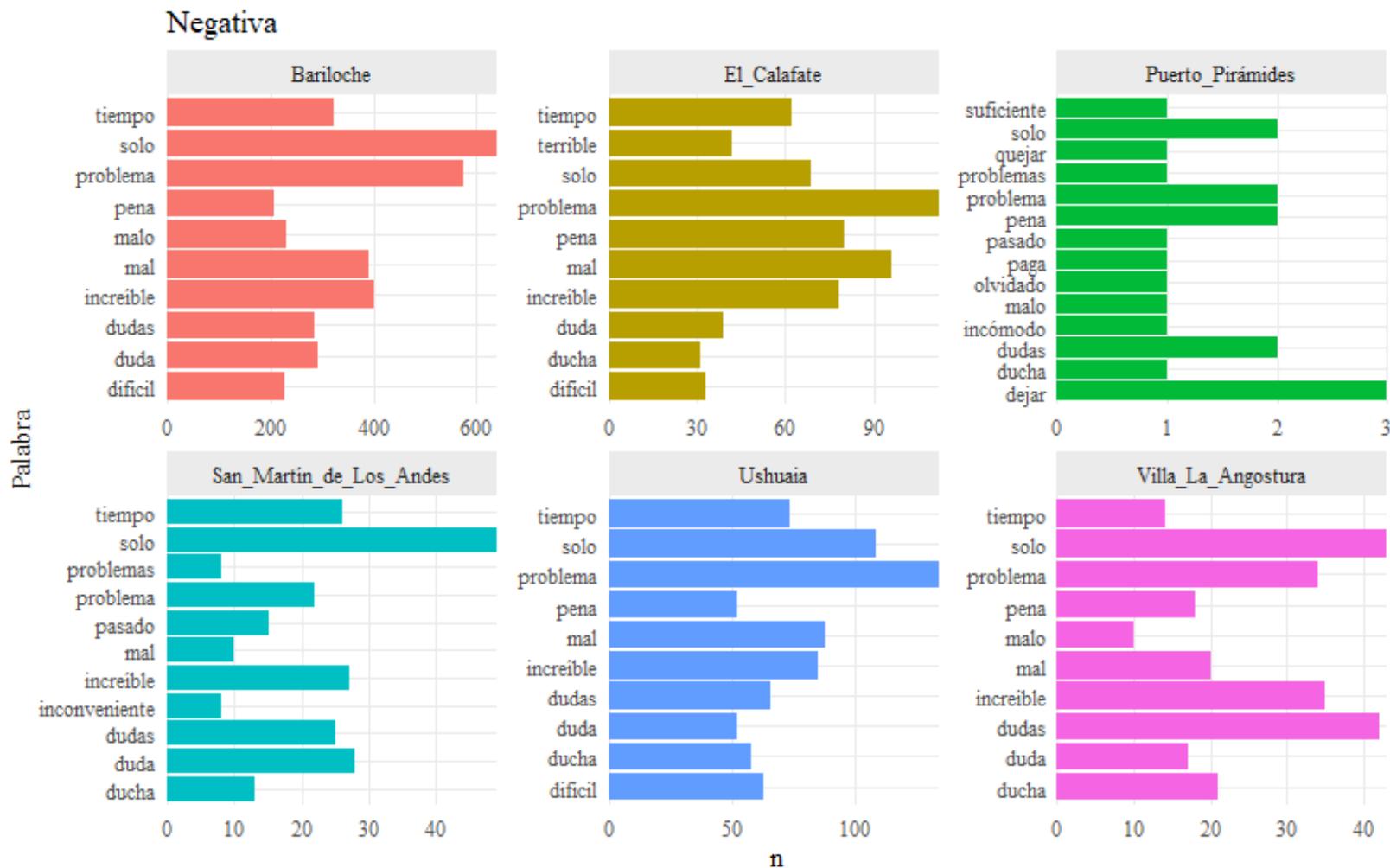
Figura 6: Opiniones positivas por ubicación del hotel



Fuente: Elaboración propia



Figura 7: Opiniones negativas por ubicación del hotel

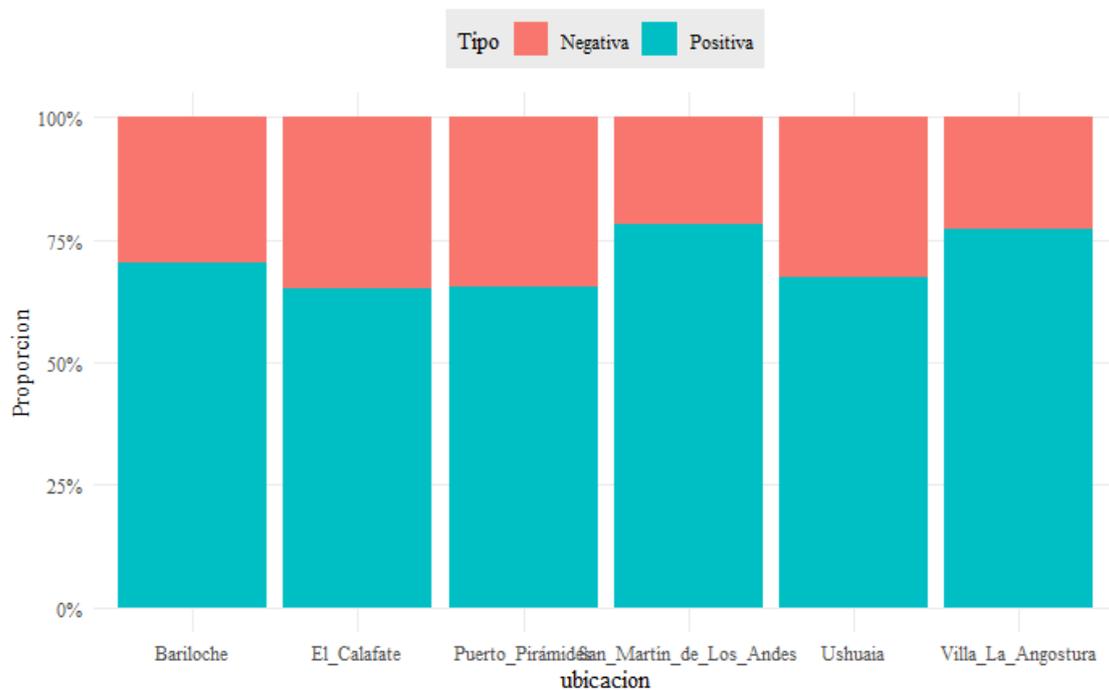


Fuente: Elaboración propia



Continuando el análisis, y discriminando también por la ubicación, la figura 8 muestra el porcentaje de opiniones positivas y negativas. En todos los casos, los guarismos indican una alta aprobación de la performance del hotel. La aprobación alcanza aproximadamente un 70% en general. En este sentido, para los administradores de los hoteles, quizás convenga centrarse en un primer momento en el contenido de las opiniones negativas para ver cuáles son los puntos para corregir y en consecuencia aplicar acciones para remediarlos. Y luego, ahondar en los comentarios positivos para potenciarlos y poder sacar el máximo provecho de ellos.

Figura 8: Proporción de opiniones según ubicación del hotel



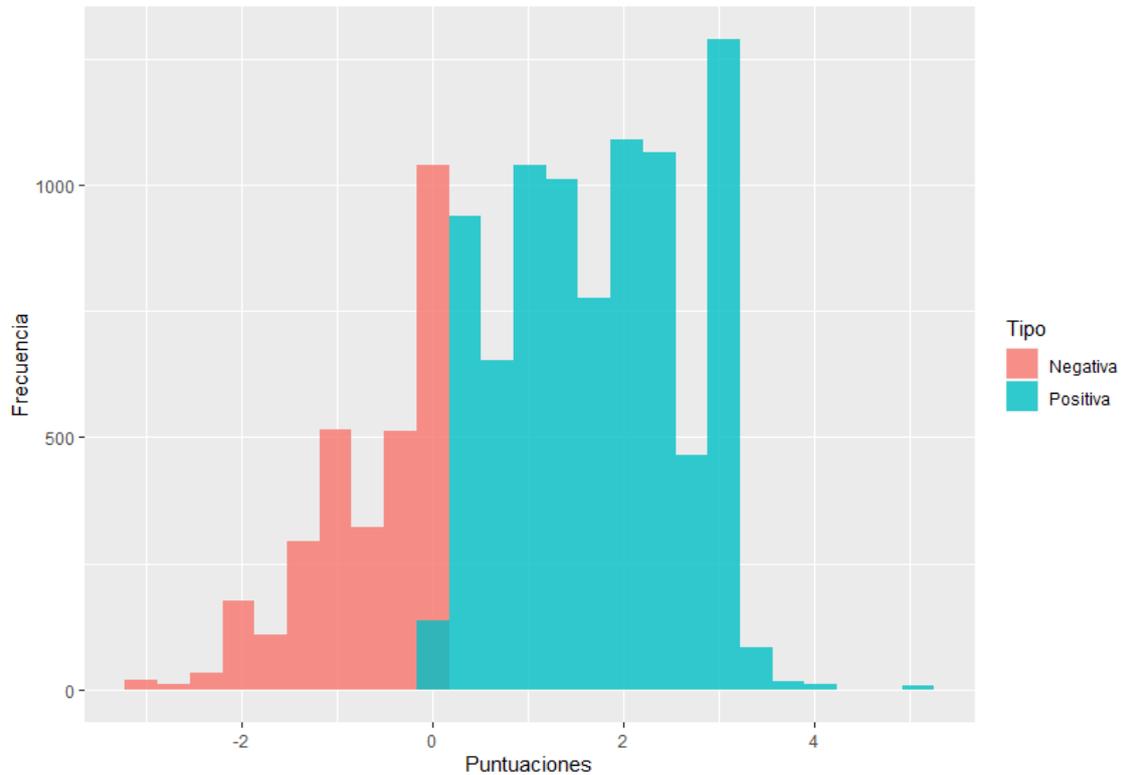
Fuente: Elaboración propia

La siguiente figura es algo similar a la anterior, pero mostrando la frecuencia de las puntuaciones sin ninguna división, sobre el total de los comentarios. Claramente las valoraciones positivas son mucho mayores a las negativas, siendo el valor 3 (tres) el que presenta una mayor frecuencia. También en este gráfico se pueden la cantidad de opiniones neutras, con valor 0 (cero), que son algo superior a las 1000 (mil).

Figura 9: Frecuencia de puntuaciones sobre el total de las opiniones

Distribución de puntuaciones

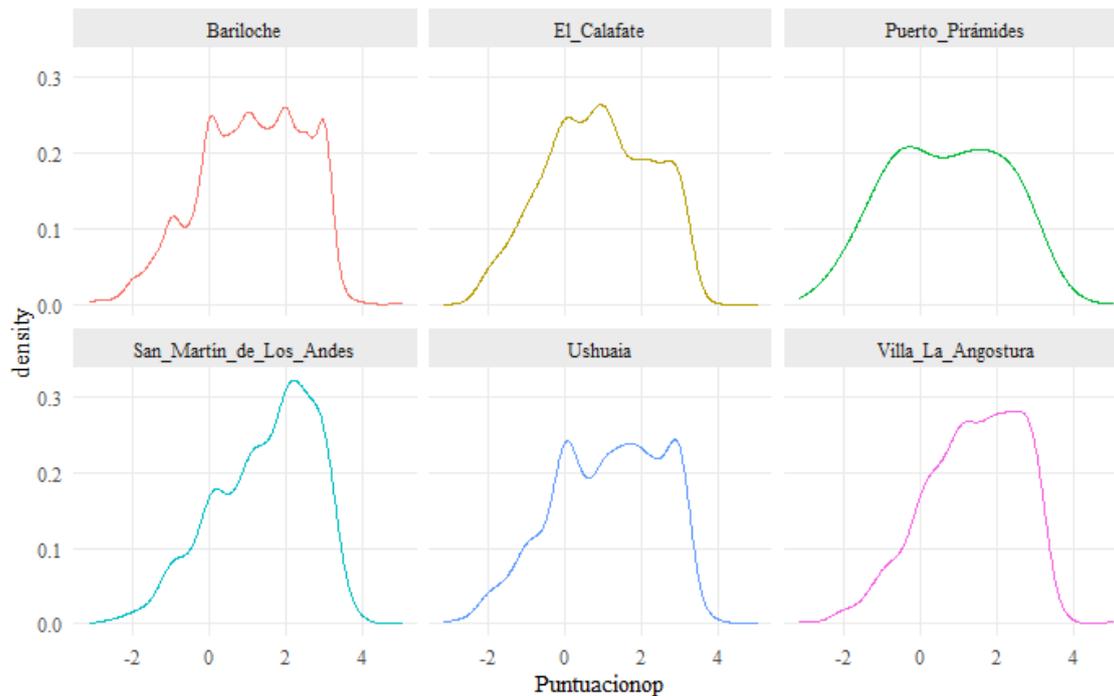
Basado en diccionario AFINN



Fuente: Elaboración propia

Para finalizar con el análisis de las puntuaciones, la figura 10 indica la distribución de los valores de cada opinión. Es decir, el porcentaje de la puntuación según la ubicación de cada hotel. En la parte inferior se encuentran las puntuaciones y en la parte izquierda los porcentajes. Siguiendo con la tendencia de comentarios positivos mayoritarios, la curva de estos gráficos tiene un desplazamiento hacia el lado, es decir mayor densidad en valores mayores a 0 (cero).

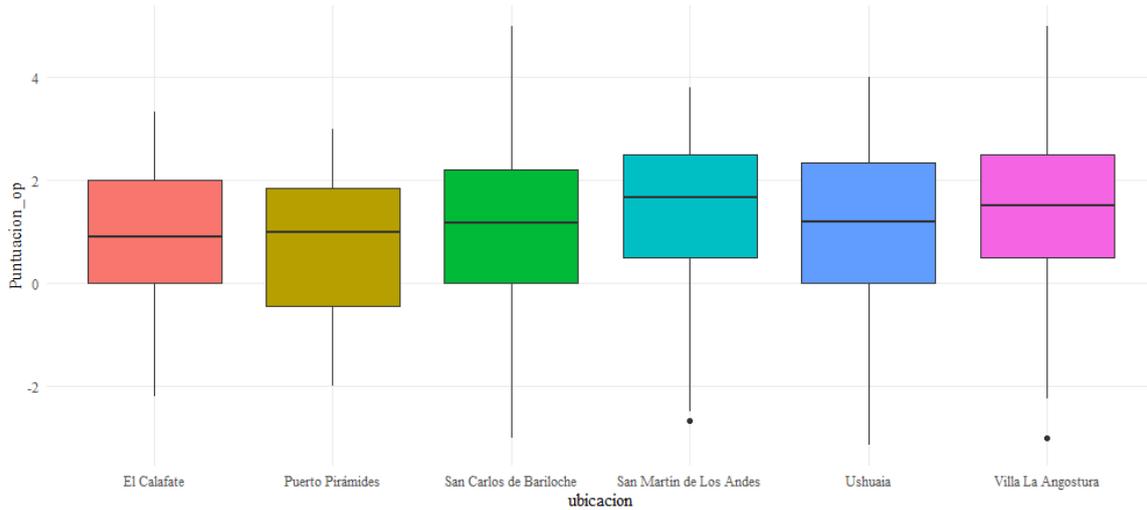
Figura 10: Distribución de opiniones según puntuaciones por ubicación



Fuente: Elaboración propia

Otra manera interesante de interpretar estos datos es mediante el diagrama de caja o bigotes (figura 11). Las líneas verticales que atraviesan cada caja representan los valores máximos y mínimos que toma la variable (en este caso, puntuación). Los puntos que pueden verse en las ubicaciones San Martín de Los Andes y Villa La Angostura, por ejemplo, son los outliers, o también llamados valores atípicos, que no están relacionados con el resto de los valores. Las líneas que se encuentran dentro de cada caja representan la mediana, o sea el valor central de la variable. El límite inferior de cada caja indica el primer cuartil, mientras que el límite superior representa el tercer cuartil. En consecuencia, cada caja agrupa el 50% de las puntuaciones, excluyendo el 25% inferior y el 25% superior.

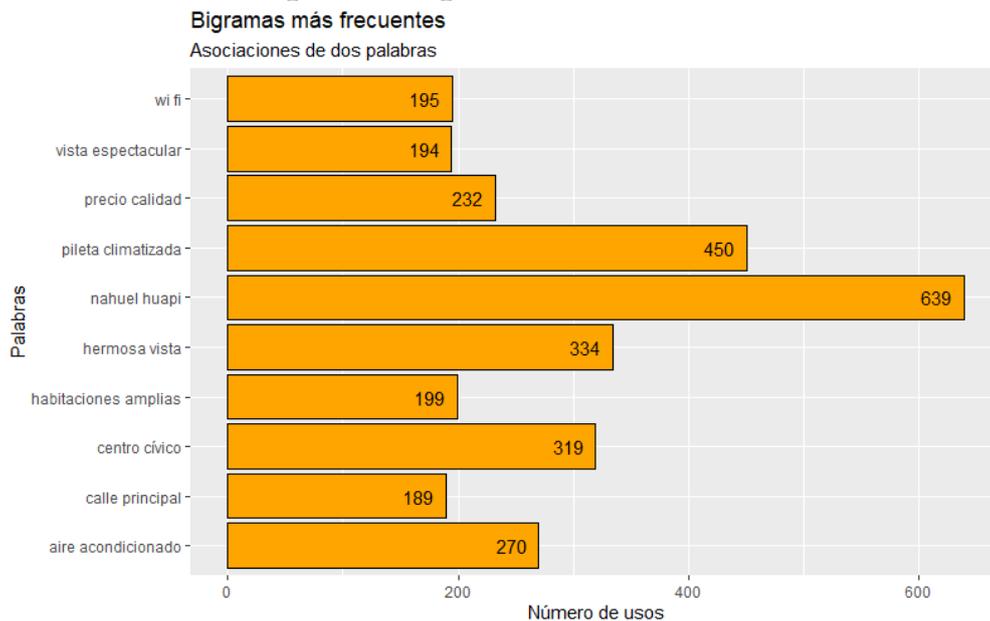
Figura 11: Diagrama de cajas o bigote



Fuente: Elaboración propia

Para finalizar, y centrándose en el análisis estricto de las palabras, se extraen las relaciones más fuertes entre ellas con el objetivo de identificar algún tipo de patrón o tendencia, esta relación se llama bi-grama. En este caso el más frecuente es *Nahuel-huapi*, le siguen *pileta-climatizada*, *hermosa-vista* y *centro-cívico*. Para los casos de *Nahuel-huapi* y *centro-cívico*, no aportan información relevante, pero para el resto se puede obtener una idea de lo que más destacan los visitantes.

Figura 12: Bigramas más frecuentes



Fuente: Elaboración propia



4. Implementación de modelos de aprendizaje automático

A continuación, se muestran distintas visualizaciones obtenidas a partir de los resultados del dataset utilizado. Los mismos fueron generados con el software PowerBi³. Asimismo, la parte final de este segmento está vinculada a las metodologías ágiles que se pueden aplicar en este proyecto, incluyendo el proceso de desarrollo de software.

4.1. Visualización

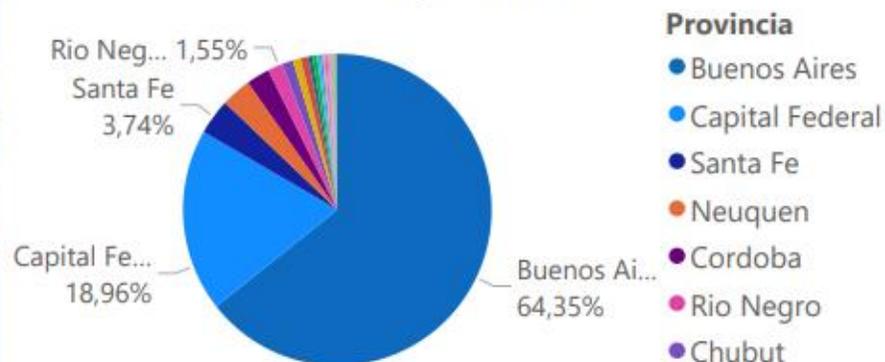
³ Software de análisis de datos que permite, mediante paneles e informes, crear visualizaciones interactivas. Es desarrollado por Microsoft.



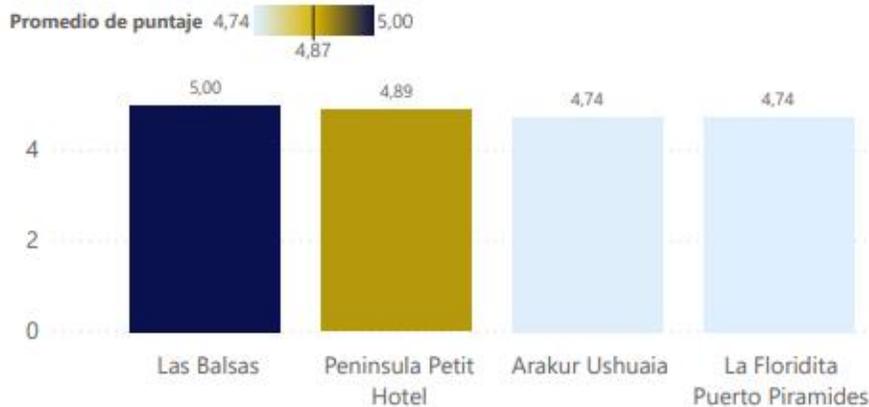
Origen de los turistas de los que se analizaron las opiniones



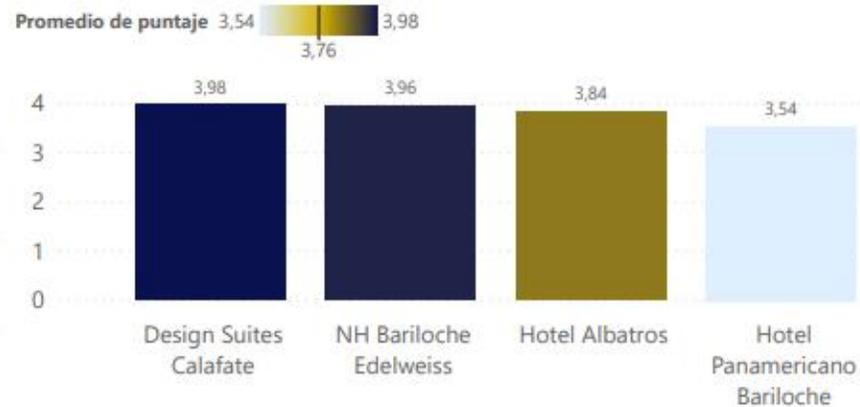
Porcentaje de opiniones por provincias argentinas



Puntaje promedio de los cuatro mejores hoteles



Puntaje promedio de los cuatro peores hoteles





4.2. Metodologías ágiles

Las organizaciones analizadas en el trabajo final de la Especialización corresponden al tipo hotelero. En este sentido, el uso de los datos resulta fundamental a la hora de obtener un “feedback” de los productos / servicios prestados, conociendo las opiniones y las valoraciones de los usuarios. Además, esta información es de acceso público, es visible para cualquier persona, y dará a conocer la reputación del hotel. En muchos casos, los potenciales clientes basarán su decisión de alojarse o no en base a estos datos. En cuanto a la arquitectura, para el caso en cuestión, la captura de datos por un lado se puede obtener mediante web scraping, ya que estamos hablando de datos no estructurados. En ciertos casos, las valoraciones pueden obtenerse mediante aplicaciones de agencias de viajes, en ese caso, estaríamos hablando de datos estructurados. Toda esta información se almacena en un data lake, que cuenta con la ventaja de agrupar en un solo lugar diferentes tipos de datos, para su posterior procesamiento y análisis. Y finalmente, el entrenamiento y el modelado se lleva a cabo mediante diferentes técnicas de Machine Learning con las que se obtienen tendencias, parámetros y demás información vinculada a la estadía de los huéspedes.

El continuo desarrollo de las tecnologías de la información convirtió ya desde hace un tiempo a las redes sociales, webs, portales, en una fuente inagotable de datos sobre lo que están pensando las personas, sus gustos, sus experiencias, etc. El sector hotelero no se encuentra excluido de este fenómeno, por lo que el uso de esta información le permitirá poder adaptarse más eficazmente al contexto en el que se encuentran.

La implementación de un modelo de aprendizaje automático permitirá identificar principalmente patrones y preferencias de consumo, algo que también se conoce con el nombre de descubrimiento de conocimiento (KDD en inglés). En el mismo se analizan grandes cantidades de datos y aplicando distintas técnicas de minería de texto, se obtienen relaciones o asociaciones que anteriormente eran desconocidas. Los resultados obtenidos deben ser válidos, novedosos, útiles y legibles. Los modelos utilizados en este aspecto tienen que ver con la frecuencia de aparición de palabras, en cuales se produce una relación (lo que se llama bi-gramas, frecuencia en la que dos palabras aparecen juntas), etc.



La metodología más acorde para este tipo de proyecto es **CRISP-DM**, ya que fue creada para proyectos de minería de datos. Consta de seis fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación. Las fases no son rígidas, pueden no seguir ese orden, y están interrelacionadas entre sí, lo que lo convierte en un proceso iterativo.

La fase de comprensión del negocio tiene que ver con los objetivos y la realización del plan de proyecto, la de comprensión de los datos se vincula con la extracción, exploración, integración y validación de los datos. La fase de preparación de los datos abarca la limpieza y selección. El modelado significa seleccionar un modelo, generarlo, probarlo y evaluarlo. La fase de evaluación implica evaluar los resultados y establecer los próximos pasos a seguir. Y, por último, la implementación está relacionada con la puesta en marcha, el monitoreo y mantenimiento del modelo.

El equipo del proyecto está conformado por los siguientes actores. En primer lugar, un data engineer, el cual se encargará de extraer e integrar la información de las diferentes fuentes y además corroborar la calidad de los datos (fase de comprensión). También existirá un data scientist (también junto a usuarios expertos) el cual llevará a cabo las tareas de análisis, procesamiento y modelado (fase de preparación). A su vez, una persona se encargará de llevar el modelo a producción y un usuario final será el que valide los resultados - QA (evaluación e implementación).



5. Conclusiones

Como vimos a lo largo de este trabajo, el proceso de obtención de datos en los hoteles se desarrolla en contextos altamente volátiles. Muchas veces, existe un cierto desfase entre su disponibilidad y el momento de toma de decisiones, por lo que para “subsanoar” esta situación se recurren a diversas técnicas como pueden ser utilizar datos actuales o datos del pasado del mismo cliente, datos anteriores de clientes similares, o la combinación de ambos. En algunos casos, los resultados brindarán una alta confiabilidad, en otros, sólo se obtendrán estimaciones o probabilidades de ocurrencia de un suceso.

En lo que concierne propiamente al análisis, las palabras que más se repiten en las opiniones analizadas son *vista*, *desayuno* y *habitaciones*, pero que sumadas representan menos del 5% de los términos. En cuanto a la connotación positiva de cada palabra, los conceptos comunes a todos los hoteles sin importar su ubicación son *bueno*, *comodidad*, *agradable*. Del mismo modo, la tendencia se repite para las palabras de connotación negativa como *problema*, *solo*, *dudas*, es decir que tanto en la positividad como en la negatividad los términos son comunes a todos los hoteles. El mayor porcentaje de distribución de las puntuaciones, según el diccionario utilizado, se sitúa en el rango entre 1 y 3, esto significa que, tomando las puntuaciones resultantes de cada opinión, la mayor cantidad se concentra entre esos valores. Para el caso de la asociación de palabras (bigramas), se obtuvieron resultados un poco triviales, como son el caso de *Nahuel-huapi* o *centro-cívico*, que no aportan información por ser el nombre de lugares. Por lo tanto, lo más relevante surge de conceptos como *pileta-climatizada*, *hermosa-vista*, *precio-calidad* o *wi-fi*, esto puede dar, al menos en principio, una idea de lo que los visitantes están buscando. Por otra parte, la base de datos utilizada contiene las opiniones de un 80% de residentes argentinos, siendo Capital Federal, Buenos Aires, Santa Fé, Neuquén y Córdoba las que concentran los mayores valores.

En resumen, se puede afirmar que existe un alto grado de aprobación sobre los servicios y la infraestructura de los hoteles de la Patagonia. Aunque existan algunas pequeñas diferencias en cuanto a los guarismos, las opiniones positivas superan el 70% en todos ellos. Llevan a cabo una gestión muy similar, esto se evidencia en las palabras vinculadas tanto para demostrar lo positivo como para lo negativo (homogeneidad).



El uso y tratamiento de todo este tipo de información está fuertemente vinculada con la privacidad de las personas. Existen diversas regulaciones en nuestro país tanto a nivel nacional como provincial que se refieren a estos aspectos. La figura de habeas data brinda la posibilidad a las personas de conocer, actualizar y rectificar los datos sobre ellas que exista en cualquier tipo de entidad, y también la de pedir la eliminación, actualización o confidencialidad de estos. Más allá de esto, en varias ocasiones no se respeta este derecho y se utiliza información sin el consentimiento de la persona, y en otras ocasiones para un propósito diferente para el cual fueron solicitados.

A futuro, se pueden definir 2 (dos) líneas de investigación: una para mejorar los resultados obtenidos, y otra que se desprende de este ensayo y servirá para ampliarlo. Con respecto a la primera, una buena medida sería mejorar la terminología del diccionario utilizado en varios análisis de este apartado. Como se dijo anteriormente, este diccionario fue tomado del idioma inglés, por lo tanto, la traducción de algunas palabras quizás no sea la más adecuada. Pero también se pueden agregar términos comunes a nuestro idioma (dialecto), lo que seguramente producirá una mayor cantidad de palabras detectadas en las opiniones (disminuirán las valoraciones neutras) y por consiguiente enriquecerá notoriamente el trabajo. Otra opción sería ampliar la base de datos utilizada, en cuanto al número de opiniones, al número de hoteles y también en cuanto a los sitios web de donde se toma la información. Sin dudas, una mayor variedad en todos estos aspectos aportará información valiosa.

La otra línea de investigación tiene que ver con la implementación de un sistema de detección automático, principalmente en redes sociales, de opiniones negativas. Esto es muy importante ya que afecta directamente en la reputación de la organización, y la acumulación o reiteración de valoraciones negativas harán dañar la imagen de la empresa. Una rápida detección de estas problemáticas permitirá actuar y, por un lado, tratar de solucionar o compensar el problema, pero también la interacción con el cliente brindará información adicional sobre una situación a corregir. En la actualidad, es un tema que cobra cada vez mayor relevancia en las organizaciones.



6. Referencias bibliográficas

1. Ampuero, L., Alfaro, R., Raymundo, C., & Dominguez, F. (2013). Modelo de madurez Tecno-organizacional para la puesta en marcha exitosa de iniciativas de Data Governance.
2. Barrera, M. C. (2014). Minería de texto: una visión actual. *Biblioteca Universitaria*, 17(2), 129-138.
3. Cano, J. L. (2007). Business Intelligence: Competir con información. En J. L. Cano, *Business Intelligence: Competir con información* (págs. 328-335). Banesto Fundación Cultural.
4. Charu, A., & ChengXiang, Z. (2012). *Mining Text Data*. Nueva York: Springer.
5. Constitución de la Nación Argentina. Artículo número 43 (1994).
6. Doyle, L. B. (1961). Semantic road maps for literature searchers. *Journal of the ACM (JACM)*, 8(4), 553-578.
7. Durán, A., Natarajan, R., & Giraldo, E. (2011). Sincronización de los ciclos de información y de gestión de infraestructuras de servicios hoteleros. *Dirección y Organización*, (45), 38-45.
8. Gantz, J., & Reinsel, D. (2012). The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. *IDC iView: IDC Analyze the future*, 2007(2012), 1-16.
9. Haddi, E., Liu, X. y Shi, Y. (2013). El papel del preprocesamiento de texto en el análisis de sentimientos. *Procedia Computer Science*, 17, 26-32.
10. Hearst, M. (s.f.). *Untangling Text Data Mining*. Berkeley, California: University of California, Berkeley.
11. Hernández, M. B., & Gómez, J. M. (2013). Aplicaciones de procesamiento de lenguaje natural. *Revista Politécnica*, 32.
12. Johnson, F. C., Paice, C. D., Black, W. J., & Neal, A. P. (1997). The application of linguistic processing to automatic abstract generation.
13. Justicia de la Torre, M. d. (2017). *Universidad de Granada*. Obtenido de <http://digibug.ugr.es/handle/10481/46975>



14. Kostoff, R. N., & DeMarco, R. A. (2001). Peer reviewed: extracting information from the literature by text mining.
15. Kruger, T. F., Acosta, A. A., Simmons, K. F., Swanson, R. J., Matta, J. F., & Oehninger, S. (1988). Predictive value of abnormal sperm morphology in vitro fertilization. *Fertility and sterility*, 49(1), 112-117.
16. Law, R., Mok, H., & Goh, C. (2007, August). Data mining in tourism demand analysis: A retrospective analysis. In *International Conference on Advanced Data Mining and Applications* (pp. 508-515). Springer, Berlin, Heidelberg.
17. Ley N°1845. Protección de datos personales (2005).
18. Ley N°25326. Protección de datos personales (2000).
19. Luck, D., & Stephenson, M. (2009). Una evaluación de los elementos significativos de la gestión de la relación con el cliente dentro de la industria hotelera. *Turismo Hoy*, 9, 7-26.
20. Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2), 159-165.
21. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute.
22. Minghetti, V. (2003). Construir el valor del cliente en la industria hotelera: hacia la definición de un sistema de información centrado en el cliente. *Tecnología de la Información y Turismo*, 6(2), 141-152.
23. Montes, M., & de Lenguaje Natural, G. L. (2001). Minería de texto: Un nuevo reto computacional. Obtenido de [https://ccc. inaoep. mx/mmontesg/publicaciones/2001/MineriaTexto-md01. pdf](https://ccc.inaoep.mx/mmontesg/publicaciones/2001/MineriaTexto-md01.pdf).
24. Nelson, D. J., Schone, P. J., & Bates, R. M. (1999). U.S. Patent No. 5,937,422. Washington, DC: U.S. Patent and Trademark Office.
25. Piller, S., Perrin, M., & Jossen, A. (2001). Methods for state-of-charge determination and their applications. *Journal of power sources*, 96(1), 113-120.
26. Sánchez Carballido, J. R. (2008). Perspectivas de la información en Internet: ciberdemocracia, redes sociales y web semántica.



27. Sanz-Magallón Delhaize, G. (2018). Oportunidades del uso del Big Data en empresas del sector de alojamiento. Implicaciones para la política de competitividad del sector turístico a partir del caso de la Unión Europea.
28. Steinmann, M., Matei, S. A., & Collmann, J. (2016). A theoretical framework for ethical reflection in big data research. In *Ethical Reasoning in Big Data* (pp. 11-27). Springer, Cham.
29. Valesani, M. E., Mariño, S. I., & Martínez, D. L. L. R. (2003). La Protección de los datos personales en los sistemas informáticos. La instrumentación en la Argentina. In *CISIC 2003: Madrid-Majadahonda, 7, 8 y 9 de mayo de 2003* (pp. 246-257).
30. Vila, T. D., & Vila, N. A. (2012). El fenómeno 2.0 en el sector turístico. El caso de Madrid 2.0. *Guía de estilo para autores*, 10, 225.
31. Vila, T. D., & Vila, N. A. (2014). Gestión de las redes sociales turísticas en la web 2.0. *Vivat Academia*, 57-78.
32. Vilares, D., Alonso, M. A., & Gómez-Rodríguez, C. (2013). Una aproximación supervisada para la minería de opiniones sobre tuits en español en base a conocimiento lingüístico. *Procesamiento del lenguaje natural*, 51, 127-134.