ISSN 1870-4069

Análisis de comentarios de turistas sobre Áreas Naturales Protegidas de México utilizando plataformas Google

Cynthia Yustis-Cruz, Magdalena Saldana-Perez, Marco Moreno-Ibarra

Instituto Politécnico Nacional,
Centro de Investigación en Computación
México

{cyustisc2020, amagdasaldana, marcomoreno}@cic.ipn.mx

Resumen. El turismo es una actividad económica de gran relevancia en México. Las Áreas Naturales Protegidas son áreas resguardadas para la conservación de su biodiversidad, en las que recientemente se ha buscado implementar estrategias que permitan aprovecharlas turísticamente de una forma sustentable. El Procesamiento de Lenguaje Natural es el campo dentro de las ciencias de la computación que conjunta técnicas de inteligencia artificial con lingüística y que, entre otras aplicaciones, permite analizar textos de forma automática. El presente trabajo es una propuesta que analiza 430 comentarios de usuarios de YouTube y Google Maps referentes al Área Natural Protegida Reserva de la Biósfera Tehuacán-Cuicatlán. El estudio consiste en un análisis de sentimientos que otorga una puntuación para identificar los comentarios positivos y negativos. Los datos fueron etiquetados manualmente y automáticamente empleando la *Natural Language API* de Google Cloud. Finalmente, se comparan los resultados obtenidos mediante el etiquetado automático con el etiquetado manual y se presentan las palabras más frecuentes en cada una de las categorías.

Palabras clave: Área natural protegida, procesamiento de lenguaje natural, análisis de sentimientos, natural language API, Google.

Analysis of Tourist Comments on Protected Natural Areas of Mexico Using Google Platforms

Abstract. Tourism is a highly significant economic activity in Mexico. Protected Natural Areas are regions designated for the conservation of biodiversity, where sustainable tourism strategies have recently been promoted. Natural Language Processing (NLP) is a field within computer science that combines artificial intelligence techniques with linguistics and enables automatic text analysis, among other applications. This study proposes the analysis of 430 user comments

from YouTube and Google Maps referring to the Protected Natural Area Tehuacán-Cuicatlán Biosphere Reserve. The analysis involves a sentiment evaluation that assigns a score to identify positive and negative comments. The data was labeled both manually and automatically using Google Cloud's Natural Language API. Finally, the results from automatic and manual labeling are compared, and the most frequent words in each sentiment category are presented.

Palabras claves: Protected natural area, natural language processing, sentiment analysis, natural language API, google.

1. Introducción

El turismo es una actividad económica de gran importancia en México [27], por ello la política sectorial federal ha mostrado interés por diversificar la oferta turística hacia sitios que permitan un mayor contacto con la naturaleza [7], como lo son las Áreas Naturales Protegidas (ANP). Por decreto, las ANP son áreas dedicadas a la protección de la biodiversidad, el cuidado de los paisajes naturales y el mantenimiento de los sistemas ecológicos que la componen, donde se busca gestionar la interacción entre pobladores y visitantes de las mismas para lograr un desarrollo social sostenible. Dichas áreas son gestionadas principalmente por la Comisión Nacional de Áreas Naturales Protegidas (CONANP) [8].
Como parte de la estrategia para el aprovechamiento responsable de las ANP se ha planteado el Marco Estratégico de Turismo Sustentable en Áreas Naturales Protegidas de México [5]. Sin embargo, no se han desarrollado las herramientas suficientes para realizar una gestión turística adecuada que permita tomar decisiones y generar nuevas propuestas basadas en un análisis de los datos que pueden proporcionar los visitantes. Es decir, no se cuenta con la información suficiente para realizar el diagnóstico sobre el comportamiento de los turistas, los sitios que visitan, su percepción de la región y la huella ecológica y digital que representan para las ANP. Es importante mencionar que, pese al avance tecnológico en otras áreas económicas, el turismo, y de manera particular las ANP carecen de las herramientas tecnológicas necesarias para la gestión y promoción eficiente de sus actividades.

Las plataformas digitales son herramientas tecnológicas que la sociedad usa día con día, para comunicarse, expresarse y promoverse. Gracias a estas acciones, los usuarios proporcionan grandes cantidades de información que son usadas por diversas compañías para diferentes objetivos [23]. Entre estos objetivos podemos mencionar la detección de preferencias de usuario, detección de patrones de movimiento en ciudades, marketing y ventas, entre muchas otras.

En el caso que aborda el presente trabajo se analizan datos generados por usuarios de YouTube y Google Maps, en los que muestran opiniones sobre sus experiencias en puntos específicos dentro del Área Natural Protegida (ANP) Reserva de la Biósfera Tehuacán-Cuicatlán [6]. El objetivo es tratar los datos empleando Procesamiento de Lenguaje Natural (PLN) para convertirlos en información que permita comprender mejor la opinión turística acerca del ANP

propuesta. Esto se logra aplicando la herramienta de análisis *Natural Language API* de Google Cloud [15] para realizar un etiquetado automático que permita clasificar los comentarios en positivos, negativos y neutros. Posteriormente, el resultado será comparado con un etiquetado manual. Finalmente, dichos resultados serán separados para entregar las palabras más relevantes de cada categoría.

La sección 2 presenta los antecedentes, que incluyen el estado del arte. Posteriormente, en la sección 3 se presenta la metodología, así como una descripción de las fuentes de datos y herramientas de análisis seleccionadas. Los resultados de los experimentos se describen en la sección 4, para ser discutidos en la sección 5 y el trabajo finaliza con la conclusión en la sección 6.

2. Antecedentes

A continuación, se presentan antecedentes relevantes para el desarrollo y comprensión del artículo. Se incluyen los antecedentes de datos turísticos y aplicaciones tecnológicas empleadas en el sector turístico, así como estado del arte relacionado con el procesamiento de lenguaje natural aplicado al análisis de sentimientos.

2.1. Datos turísticos en México

La CONANP y el gobierno de México cuentan con sitios web [11] que permiten acceder a información relacionada con cada una de las ANP de México. Sin embargo, esta información es muy general y no es suficiente para la gestión turística de dichas áreas. Por un lado, se cuenta con fichas técnicas [4] que permiten conocer información como la categoría que las áreas tienen asignada, la macro ubicación de la misma, extensión territorial, número de habitantes, datos relacionados con el decreto oficial y personal encargado de dirigirla. También se cuenta con un sistema de información geoespacial [9] que proporciona mapas de las delimitaciones territoriales de las ANP, así como mapas disponibles para ser descargados en formatos adecuados para su edición en programas para el procesamiento de datos geoespaciales. Así mismo, existen otras plataformas gubernamentales como es el sitio web de biodiversidad de la Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO) [3] que también proporciona información general de ANP.

Por su parte, la Secretaria de Turismo (SECTUR), cuenta con el sistema DataTur [24], una plataforma que proporciona información sobre vuelos, visitas a museos y zonas arqueológicas, indicadores de empleo turístico, encuestas a viajeros internacionales, entre otros datos. De igual forma, proporciona datos geoespaciales referentes a mapas de algunas entidades y rutas turísticas, se puede encontrar una recopilación de dichos datos en [30]. También, cuenta con información sobre atractivos y destinos turísticos dirigida a los visitantes [28]. No obstante, estos datos son dirigidos, en su mayoría, a turistas y no a gestores, además de que hacen referencia al turismo general en México y entre sus estadísticas no se encontró información sobre los turistas que visitan las ANP.

2.2. Tecnologías de la información aplicadas a la gestión del turismo en México

Se denomina Turismo Inteligente al uso de tecnologías de la información, tales como Internet de las Cosas (IoT), Cómputo en la Nube (*Cloud Computing*), tecnologías móviles e Inteligencia Artificial (AI) para el desarrollo de sistemas inteligentes aplicados al turismo, que a su vez son amigables con los usuarios. En [20] se menciona que el futuro del turismo está basado en el desarrollo tecnológico, la rapidez y la extensión de la aplicación de dichos sistemas a todos los niveles del sector turístico. Se ha estudiado y concluido que se debería tener conocimiento de las áreas que resultan más atractivas para los visitantes y cuáles son sus características principales. La importancia de esta información radica en que estos datos contribuyen como herramientas para un mejor y más inteligente diseño de propuestas, así como para la implementación de estrategias sustentables que llevarán a la optimización de la experiencia turística.

Desde el año 2013, la Sociedad Estatal para la Gestión de la Innovación y las Tecnologías Turísticas (SEGITTUR), una sociedad española dedicada a la gestión e innovación de tecnologías turísticas, junto con la Secretaria de Estado de Turismo de España han impulsado un programa de Destinos Turísticos Inteligentes (DTI) [29]. En 2015, la isla de Cozumel se incorporó a este proyecto, no obstante, el gobierno de México remarcó la necesidad de trabajar en la infraestructura para sistemas de comunicaciones y desarrollo de plataformas tecnológicas, que permita la interacción del gobierno con los prestadores de servicios para establecer nuevos modelos de negocios y operación [26] [25].
Por otro lado, la Zona Dorada de Ciudad Juárez, ubicada en el norte de México, fue tomada como caso de estudio para la búsqueda de áreas de oportunidad hacia la conformación de la zona como un espacio turístico inteligente y sustentable, lo anterior, mediante un estudio de sus componentes e infraestructura tecnológica. Derivado del análisis se concluyó que la zona estudiada no cuenta con el soporte tecnológico para clasificar como DTI, siendo una de las principales limitantes el acceso a redes de internet inalámbrico abiertas. También se mencionó la falta de aplicaciones móviles para facilitar la interacción con los visitantes y un déficit en la recolección de datos que los usuarios pueden estar generando y que podrían ser aprovechados para la creación de un observatorio turístico [12].

Con el desarrollo de este trabajo se pretende estudiar los datos que pueden proporcionar las plataformas digitales y generar información de interés sobre la opinión de los turistas o potenciales turistas del Área Natural Protegida seleccionada. Así como observar el comportamiento de los datos al clasificarlos manualmente y automáticamente. Lo anterior, mediante una metodología propuesta que sea aplicable al estudio de otras Áreas Naturales Protegidas en México.

La investigación se presenta como una propuesta para aprovechar la información generada por usuarios de plataformas digitales en la búsqueda de mejorar la gestión turística en México. Se propone la incorporación de una herramienta de PLN para conocer la opinión de los turistas en un Área Natural Protegida con la visión de proporcionar alternativas que apoyen la

toma de decisiones de administradores del sector turístico. Contar con una herramienta tecnológica que permita generar información a partir del análisis de datos referentes al comportamiento y opiniones de los turistas en ANP representa un avance de impacto en materia de gestión turística a través de la integración de tecnologías de la información como una propuesta para mejorar la administración y promoción de las mismas.

2.3. Procesamiento de lenguaje natural y análisis de sentimientos

Se conoce como Procesamiento de Lenguaje Natural (PLN) al área dentro de las ciencias de la computación que se ocupa de trabajar con aspectos relacionados a la comprensión de la comunicación humana [18]. Esto puede incluir tanto la capacidad de las máquinas para comprender el lenguaje, como la capacidad de generar respuestas basadas en el mismo. Para el desarrollo de algoritmos de PLN se suele tomar un enfoque lingüístico, es decir, se toman elementos semánticos y sintácticos básicos como características [18]. También, se han implementado técnicas de aprendizaje automático para realizar tareas de clasificación dentro del área.

El análisis de sentimientos es una de las áreas más activas dentro del PLN. Su tarea es evaluar lenguaje escrito o hablado y determinar si las expresiones evaluadas son negativas, positivas o neutras [18]. Este tipo de análisis es útil en temas relacionados con atención a cliente u opinión de usuarios, medición de estado de ánimo y rastreo de comportamiento humano. Sin embargo, tanto el PLN en general como el análisis de sentimientos son temas complejos que son motivo de libros y artículos totalmente dedicados al tema [1] [2].

Actualmente, existen diversas alternativas para realizar PLN y se pueden ejecutar algoritmos de análisis en distintos lenguajes de programación. Dependiendo de las herramientas que se elijan, el procesamiento puede ir desde la limpieza de los datos hasta la segmentación del texto para el entrenamiento de modelos de clasificación, o bien, existen modelos previamente entrenados que analizan textos en diferentes idiomas. Dichos modelos se pueden encontrar en bibliotecas o módulos previamente desarrollados para distintos lenguajes de programación, o bien, dentro de plataformas de servicios de cómputo en la nube como son Microsoft Azure, Amazon Web Services o Google Cloud. En el caso que se aborda en este trabajo se utilizará la *Natural Language API* de Google Cloud [15].

Un ejemplo del análisis de sentimientos aplicado al turismo puede ser [19], donde se emplea una metodología que emite evaluaciones positivas, negativas y neutrales a datos turísticos tomados de las plataformas TripAdvisor y VirtualTourist. Concluyen que sus aportes pueden ser útiles a los encargados de gestión de los destinos turísticos en su tarea de mejorar la calidad de los servicios. También mencionan la necesidad de profundizar en la generación de datos para la posterior aplicación de ciencia de datos para su análisis.

En 2020, [21] presenta un análisis de sentimientos aplicado a una comunidad de estudiantes, esto se realizó analizando 2866 publicaciones y 3630 comentarios en inglés, obtenidos de la plataforma Google Plus (G+). Para ello hicieron uso de

Fig. 1. Metodología propuesta. (Iconos tomados de [10] y [21])

la *Natural Language API* de Google Cloud. Aunque este trabajo está totalmente dirigido al sector educativo, pudieron implementar de forma satisfactoria la herramienta de PLN de Google Cloud para el análisis, aunque destacaron que, desafortunadamente, uno de los idiomas oficiales de la región de estudio no estaba incluido en la lista de idiomas de la *API* y ello los llevó a limitar su base de datos.

3. Materiales y métodos

El presente trabajo es una propuesta de aplicación de procesamiento de lenguaje natural aplicado a comentarios de usuarios de las plataformas YouTube y Google Maps. Se estudia el caso de la Reserva de la Biosfera Tehuacán-Cuicatlán. En esta sección se describe la metodología establecida, la descripción de las fuentes de datos y la obtención de los mismos, así como la herramienta de análisis automático seleccionada.

3.1. Metodología

La Fig. 1 muestra un diagrama de la metodología establecida para este trabajo, consta de 14 pasos que van desde la selección de las plataformas para la obtención de los datos, hasta la comparación de los resultados obtenidos.

Las plataformas seleccionas para obtener los datos fueron la red social YouTube y la plataforma Google Maps. Como se muestra en el paso 2, los datos fueron obtenidos a través de las Interfaces de Programación de Aplicaciones

(*API*) oficiales de cada una de las plataformas y posteriormente se conjuntaron en un solo archivo durante el paso 3. Una vez que se tuvo un primer conjunto de datos se filtraron y etiquetaron manualmente para dejar listo el conjunto de datos que contiene información referente a la reserva natural de interés. Se eliminaron los datos que no estaban relacionados con el área natural seleccionada, dando un total de 430 comentarios relacionados, a estos datos finales se les otorgó una puntuación de -1 a 1, donde los valores cercanos a 1 representan comentarios que se relacionaron con un sentimiento positivo, los cercanos a -1 con sentimientos negativos y los cercanos a 0 se consideran comentarios neutros.

Una vez que se cuenta con el conjunto de datos de interés, y antes de realizar el análisis de sentimientos con la herramienta seleccionada, es necesaria una limpieza básica del texto donde se eliminan símbolos como emoticones, hipervínculos y otros caracteres especiales que pudiera contener el texto. Después del paso 6 se tiene un texto plano en minúsculas, listo para entrar a la herramienta de clasificación, en el paso 7 se implementa la herramienta de procesamiento de lenguaje natural de Google Cloud a través de la *Natural Language API*. Por medio de este último paso se le otorga otra etiqueta a los datos, esta etiqueta también va de -1 a 1, pasando por 0, para comentarios negativos, positivos y neutros respectivamente. Se realiza la diferencia de los valores absolutos del etiquetado manual con el etiquetado automático para saber qué tan distantes son unos valores de otros y se toma esto como el cálculo del error del paso 9.

Una vez que los datos tienen las etiquetas necesarias las frases son segmentadas en palabras durante el paso 10. Una vez en el paso 11, se eliminan las palabras conocidas como palabras vacías o *stopwords*, que incluyen clases de palabras como artículos, pronombres y preposiciones.

A través de las etiquetas generadas durante los pasos 4 y 8, durante el paso 12 se categorizan los datos en tres, los datos que tienen valores entre -1 y -0.2 se catalogan como negativos, los que tienen valor entre 1 y 0.2 como positivos y los restantes como neutros. Esta categorización se realiza tanto para el etiquetado manual como para el automático. De esta forma se tienen las categorías: manual positivo, manual negativo, manual neutro, automático positivo, automático negativo y automático neutro. Finalmente, en el paso 13 se realiza un conteo de las palabras que se repiten más en cada una de las categorías y se comparan los resultados de ambos etiquetados durante el paso 14.

A continuación, se presenta la descripción detallada de algunos de los pasos de esta metodología para su mejor comprensión. Esto incluye la descripción detallada de la obtención de datos de Youtube y el uso de la *API* para su descarga. De la misma forma para Google Maps y detalles sobre el uso de la *API* de procesamiento de lenguaje natural de Google Cloud.

3.2. Datos de Youtube

La plataforma YouTube es una red social de Google basada en publicación de videos [31], en los cuales usuarios pueden realizar comentarios. Los datos

de YouTube utilizados para este trabajo consisten en la lista de comentarios provenientes de los primeros 50 videos relacionados a una búsqueda.

Para la realización del presente proyecto, se determinó una lista de palabras para búsqueda y se realizó la obtención de los comentarios mediante la Interfaz de Programación de Aplicaciones (*API*) *YouTube Data API* oficial [17], utilizando el lenguaje de programación Python [22]. La lista de búsquedas incluye nombres de zonas y comunidades de interés dentro del ANP, mismos que se presentan a continuación:

- Tehuacán Cuicatlán
- Salinas Zapotitlán
- Alfareros Mezontla
- Jardín botánico Helia Bravo Hollis Zapotitlán

3.3. Datos de Google Maps

La plataforma de Google Maps es una aplicación de mapas web de Google [14], que permite a los usuarios realizar búsqueda de localizaciones, servicios, trazado de rutas, vistas de mapas desplazables e imágenes satelitales. Google Maps permite dar de alta ubicaciones de servicios, negocios, parques, escuelas, entre otros, y permite a los usuarios y visitantes dejar reseñas y comentarios sobre las ubicaciones.

Para la tarea de recolección de esta propuesta, también se realizó una lista de palabras para búsqueda y se realizó la obtención de los comentarios mediante la *Places API* oficial [16], utilizando el lenguaje de programación Python [22]. A diferencia de YouTube, la obtención de reseñas de lugares en Google Maps tienen la limitante de que solo permite la obtención de hasta cinco reseñas por lugar, a menos de que el solicitante sea el administrador de la ubicación. La lista de búsquedas incluye nombres de zonas y comunidades de interés dentro del ANP, mismos que se presentan a continuación:

- Cañón Alas Verdes
- Tierra Colorada
- Museo Comunitario Paleontológico de San Juan Raya
- Turritelas
- Huellas de Dinosaurios
- Reserva de la Biósfera Tehuacán-Cuicatlán Zapotitlán Salinas, Puebla
- Zapotitlan Salinas Botanical Garden
- Helia Bravo Hollis Botanical Garden
- Letras Zapotitlán
- Pata de Elefante
- Paleoparque Las ventas
- Mirador ecológico San Antonio Texcala
- El Encinal, Santa Ana Teloxtoc
- Los Reyes Metzontla, Puebla
- Centro Artesanal Comunitario - Alfareros Popolocas de los Reyes Metzontla

- Las Salinas Grandes
- Mirador Tehuacán-Oaxaca
- San Pedro Jaltepetongo
- Reserva de la Biósfera Tehuacán- Cuicatlán Cerro Prieto, Oaxaca
- Temazcal Zapotitlán
- Campestre Zapotitlán

La lista de búsqueda para lugares en Google Maps es mayor a la lista de búsqueda de YouTube debido a que las búsquedas en YouTube regresan videos relacionados y se tendrían muchos datos repetidos al buscar todos los lugares en la última lista presentada, pero Google Maps devuelve los comentarios de lugares específicos por lo que no se tendrán comentarios repetidos, a menos que el usuario haya escrito el mismo comentario en dos ubicaciones diferentes.

3.4. Google Cloud para el procesamiento del lenguaje natural

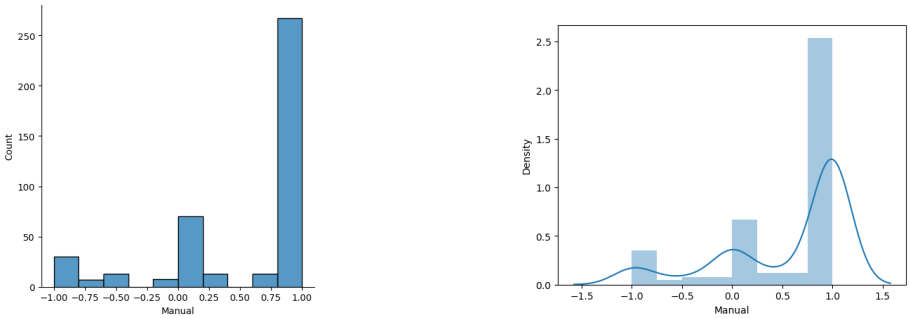
Google Cloud es un conjunto de servicios de cómputo en la nube ofertados por Google [13]. Como parte de estos servicios, éxiste la *Natural Language API* que utiliza técnicas de aprendizaje automático para analizar textos no estructurados. Esta *API* cuenta con modelos previamente entrenados que permiten realizar funciones de análisis de opinión, análisis de entidades, análisis de opinión por entidad, clasificación de contenido y análisis de sintaxis. Es importante mencionar que la *Natural Language API* es solamente uno de los tres servicios de procesamiento de lenguaje natural que ofrece Google. Se eligió para probar su capacidad de clasificar la opinión de los textos con los modelos previamente entrenados.

La documentación de Google Cloud para el procesamiento de lenguaje natural describe el análisis de sentimientos como una inspección de textos que identifica la opinión emocional que prevalece dentro del mismo, lo anterior para determinar la actitud positiva, negativa o neutra del autor. Como resultado de la solicitud la *API* entrega una puntuación que va de -1 a 1, donde valores cercanos a -1 corresponden a sentimientos negativos, los cercanos a 1 a sentimientos positivos y los valores alrededor de 0 a comentarios neutros. El método que permite efectuar esta tarea puede analizar un texto general o a nivel de oraciones. Para la metodología propuesta se utilizó la puntuación obtenida a nivel texto.

4. Experimentos y resultados

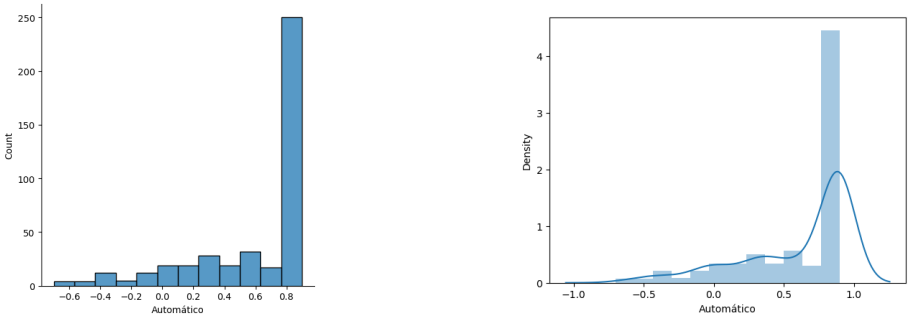
En esta sección se presentan las gráficas obtenidas siguiendo la metodología establecida, así como los resultados y experimentos adicionales.

Las Fig. 2 y 3 muestran la distribución de los valores obtenidos en el etiquetado manual y el automático. A simple vista se observa que la distribución es muy similar para valores cercanos a 1, es decir, en los valores referentes a la comentarios positivos. Esta distribución va mostrando más cambios para valores menores a 0.5, es decir, para hacia los valores relacionados con los comentarios neutros y negativos.



(a) Distribución **(b)** Densidad

Fig. 2. Distribución de los valores otorgados mediante el etiquetado manual. Se representa la distribución de la puntuación, entre -1 a 1, de cada uno de los 430 comentarios.



(a) Distribución **(b)** Densidad

Fig. 3. Distribución de los valores otorgados mediante el etiquetado automático. Se representa la distribución de la puntuación, entre -1 a 1, de cada uno de los 430 comentarios.

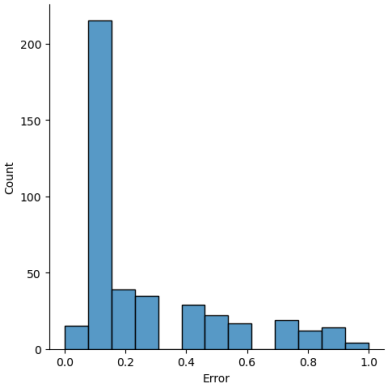


Fig. 4. Distribución del error calculado a través de obtener la diferencia entre la puntuación asignada de manera manual y automática.

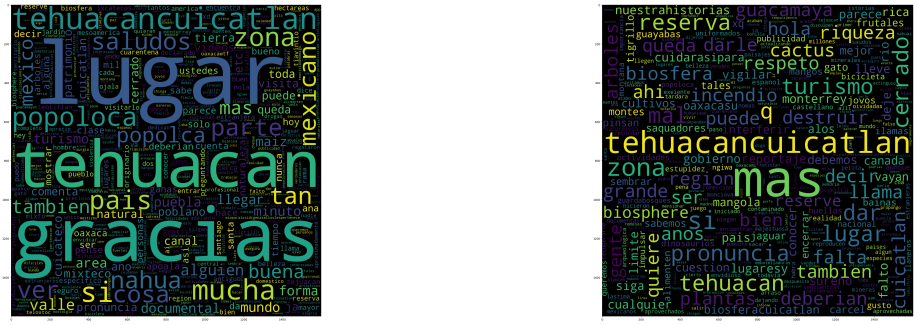
(a) Manual **(b)** Automático

Fig. 5. Nube de palabras de la categoría positiva para el etiquetado manual y automático.

La Fig. 4 muestra la distribución obtenida del error calculado, este error se obtuvo del valor absoluto de la diferencia entre valores absolutos del etiquetado manual y el etiquetado automático. Aunque se muestra que la mayoría de los errores calculados fueron cercanos a 0.1, al obtener el promedio de los valores se tiene que en promedio se tuvo un error de 0.26. Es importante mencionar que la clasificación manual se realizó basada en listas de términos establecidos para diferenciar entre positivos y negativos. Los datos que no tuvieran alguno de ellos se determinaron como neutros. Si se tenían varias frases y cada una expresaba diferentes sentimientos se tomaba el promedio de las puntuaciones. La clasificación manual se logró estableciendo la semántica entre los grupos de

Pos	Manual	Automático
1	('lugar')	('lugar')
2	('tehuacan')	('tehuacan')
3	('mas')	('gracias')
4	('gracias')	('mas')
5	('hermoso')	('hermoso')
6	('excelente')	('excelente')
7	('reserva')	('reserva')
8	('mexico')	('mexico')
9	('bonito')	('puebla')
10	('puebla')	('ver')
11	('tambien')	('tambien')
12	('ver')	('si')
13	('conocer')	('saludos')
14	('gran')	('bonito')
15	('naturaleza')	('tan')
16	('saludos')	('conocer')
17	('san')	('pais')
18	('lugares')	('san')
19	('cuicatlan')	('lugares')
20	('bien')	('zona')
21	('si')	('bien')
22	('zona')	('naturaleza')
23	('ser')	('ser')
24	('oaxaca')	('gran')
25	('mejor')	('oaxaca')
26	('tan')	('cuicatlan')
27	('aqui')	('muchas')
28	('ir')	('parte')
29	('recomiendo')	('documental')
30	('region')	('buen')

Fig. 6. Comparación de las palabras más frecuentes en la categoría positiva para el etiquetado manual y automático.

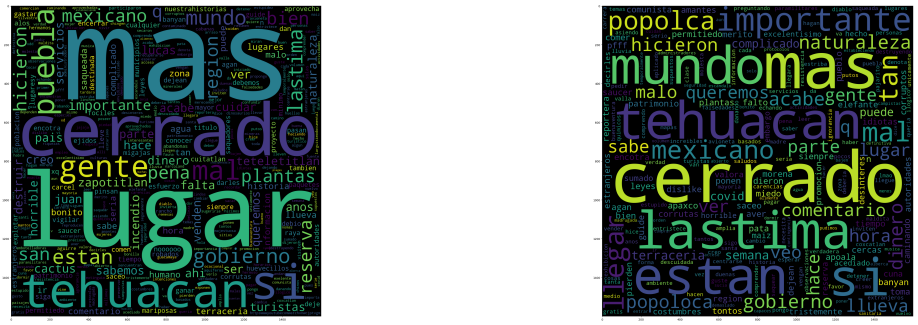


(a) Manual **(b)** Automático

Fig. 7. Nube de palabras de la categoría neutra para el etiquetado manual y automático.

palabras que se referían a aspectos positivos y negativos. Entre estas palabras no se incluyen los nombres propios de los sitios de interés.

Cuando los comentarios contaban con ambas etiquetas fueron separados en subgrupos. Una vez categorizados se utilizó la herramienta nube de palabras para generar una visualización de las palabras más frecuentes en cada categoría. El tipo de gráfico presentado en la Fig. 5 muestra las palabras más frecuentes empleadas en los textos de la categoría positiva, es decir, comentarios con valores



(a) Manual **(b)** Automático

Fig. 8. Nube de palabras de la categoría negativo para el etiquetado manual y automático.

de etiquetado en el intervalo que va de 0.2 a 1, para el método manual en la Fig. 5a y el automático en la Fig. 5b. El tamaño de las palabras mostradas es proporcional a la cantidad de veces que se repiten en la categoría, las palabras con mayor tamaño han sido mencionadas más frecuentemente que las de menor tamaño. Las cercanía de unas palabras con otras también está relacionada con su uso en el texto, es decir, las palabras más cercanas son las que están comúnmente agrupadas en el texto.

En la Fig. 6 se presentan las 30 palabras más frecuentes en la categoría positiva para los comentarios etiquetados manualmente y automáticamente. La primera posición de la lista es la palabra más repetida y descienden conforme desciende su frecuencia de uso. Se usa un semáforo de colores, donde el color verde representa que la palabra se encuentra en ambas listas en la misma posición de lista, el amarillo indica que la palabra se encuentra en ambas listas pero están desfasadas una posición, mientras que el naranja representa que se encuentra en ambas listas pero desfasadas en más de una posición y el rojo representa que la palabra no se encuentran en la otra lista.

Se observa que, de las 30 palabras comparadas solo cinco palabras de cada lista son diferentes y hubo coincidencia en 25 palabras de uso frecuente. Entre las palabras más frecuentes encontramos: lugar, Tehuacán, hermoso, excelente, reserva, México, bonito, Puebla, Naturaleza, conocer y país.

Estos ejercicios se repitieron para la categoría de negativos y neutros. En la Fig. 7 podemos observar la nube de palabras para la categoría de comentarios neutros, es decir, con valores de etiquetado mayores a -0.2 y menores a 0.2. Por su parte, las nubes de palabras para la categoría de comentarios negativos, con un etiquetado entre -0.2 y -1, se pude observar en la Fig. 8.

La comparación de las palabras más frecuentes para el etiquetado manual y automático de las categorías de comentarios neutros y comentarios negativos se presenta en la Fig. 9. La Fig. 9a muestra las palabras de la categoría de comentarios neutros y la Fig. 9b las palabras de la categoría de comentarios

Pos	Automático	Manual
1	('mas')	('lugar')
2	('tehuacancuicatlan')	('gracias')
3	('si')	('tehuacan')
4	('mal')	('tehuacancuicatlan')
5	('zona')	('si')
6	('q')	('muchas')
7	('turismo')	('pais')
8	('pronuncia')	('tan')
9	('reserva')	('parte')
10	('tehuacan')	('ver')
11	('lugar')	('saludos')
12	('the')	('zona')
13	('dar')	('popolca')
14	('cerrado')	('cosas')
15	('bien')	('tambien')
16	('biosfera')	('mexico')
17	('deberian')	('mas')
18	('arboles')	('alguien')
19	('quiere')	('nahuas')
20	('decir')	('valle')
21	('plantas')	('llegar')
22	('cactus')	('popolocas')
23	('gente')	('puebla')
24	('ser')	('mundo')
25	('falta')	('maiz')
26	('destruir')	('area')
27	('ahi')	('mexicano')
28	('incendio')	('documental')
29	('anos')	('buena')
30	('tambien')	('forma')

(a) Categoría de comentarios neutros.

Pos	Manual	Automático
1	('mas')	('cerrado')
2	('lugar')	('mas')
3	('cerrado')	('si')
4	('tehuacan')	('tehuacan')
5	('si')	('lastima')
6	('gente')	('mundo')
7	('puebla')	('estan')
8	('mal')	('llegar')
9	('gobierno')	('tan')
10	('estan')	('mal')
11	('mundo')	('popolca')
12	('lastima')	('gobierno')
13	('plantas')	('acabe')
14	('bien')	('dinero')
15	('reserva')	('naturaleza')
16	('region',)	('hicieron')
17	('san')	('queremos')
18	('q')	('ver')
19	('hicieron')	('hace')
20	('pena')	('parte')
21	('mexicanos')	('lugar')
22	('pais')	('importante')
23	('cuidar')	('veo')
24	('ser')	('malo')
25	('zapotitlan')	('sabe')
26	('venden')	('q')
27	('turistas')	('llueva')
28	('sabemos')	('gente')
29	('parte')	('hora')
30	('juan')	('terraceria')

(b) Categoría de comentarios negativos.

Fig. 9. Comparación de las palabras más frecuentes en la categoría neutra y negativa para el etiquetado manual y automático. El color verde representa que la palabra se encuentra en la ambas listas con la misma posición, el amarillo indica que la palabra se encuentra en ambas listas pero desfasadas una posición, el naranja representa que se encuentra en ambas listas pero desfasadas en más de una posición y el rojo representa que las palabras no se encuentran en la otra lista.

negativos. El semáforo de colores, explicado previamente, se mantiene para estas figuras. Con esto podemos visualizar que la cantidad de palabras frecuentes que se repiten es más reducida que en el caso de la categoría de comentarios positivos, siendo la categoría de comentarios neutros la que tiene menor cantidad de palabras repetidas en común entre el etiquetado manual y automático.

Sin embargo, en la Fig. 9 podemos observar que hay palabras repetidas que comparten ambas categorías, es por esto que se propuso un último ejercicio que consiste en unir la categoría de comentarios neutros con negativos para evaluar si existen más palabras repetidas en común. Se creó una categoría de prueba que va de 0.2 a -1 y se realizó el ejercicio de conteo, los resultados de éste pueden ser visualizados en la Fig. 10.

Pos	Manual	Automático
1	('mas')	('lugar')
2	('cerrado')	('mas')
3	('si')	('tehuacan')
4	('tehuacan')	('si')
5	('tehuacancuicatlan')	('cerrado')
6	('mal')	('gracias')
7	('q')	('pais')
8	('lugar')	('puebla')
9	('zona')	('tehuacancuicatlan')
10	('reserva')	('tan')
11	('mundo')	('muchas')
12	('gente')	('mundo')
13	('lastima')	('parte')
14	('plantas')	('ver')
15	('gobierno')	('zona')
16	('bien')	('saludos')
17	('popolca')	('mal')
18	('turismo')	('bien')
19	('naturaleza')	('cosas')
20	('queremos')	('reserva')
21	('pronuncia')	('gente')
22	('hicieron')	('tambien')
23	('anos')	('region')
24	('region')	('estan')
25	('estan')	('llegar')
26	('puede')	('mexicanos')
27	('the')	('mexicano')
28	('dar')	('gobierno')
29	('falto')	('popolca')
30	('llegar')	('ser')

Fig. 10. Comparación de las palabras más frecuentes en la categoría neutro-negativo propuesta para el etiquetado manual y automático.

En la Fig. 10 podemos observar que la cantidad de palabras repetidas que coinciden para ambos etiquetados aumentó en comparación con la Fig. 9, teniendo en total 17 palabras de alta frecuencia de repetición para ambos etiquetados. Siendo algunas de ellas: más, cerrado, Tehuacán, tehuacancuicatlan, gobierno, región, zona, llegar, reserva y mundo.

Con este último ejercicio se concluyen los experimentos realizados para este trabajo, en la siguiente sección se discuten los resultados obtenidos.

5. Discusión

En este apartado se comentarán los resultados de la sección anterior, así como algunas propuestas de mejora y trabajo a futuro.

Comenzaremos mencionando que el filtrado de los datos para obtener un cuerpo de comentarios enfocado en el turismo del ANP seleccionada se realizó manualmente ya que aún no se contaba con suficientes datos para realizar un

modelo de clasificación automática. Sin embargo, es una tarea que se plantea para ser desarrollada en futuros trabajos.

En las gráficas de distribución presentadas en las Fig. 2 y 3 se puede observar una distribución similar, donde la mayoría de los comentarios fueron etiquetados con valores correspondientes a sentimientos positivos y los comentarios neutros y negativos representaban la minoría de los datos. Al mostrar el error que se tiene, se encuentra que la variación promedio entre el etiquetado manual y el automático es de aproximadamente 0.26. Para este caso es importante tomar en cuenta que existe un factor de objetividad en el etiquetado manual de los datos, pues la puntuación es ponderada mediante un listado de términos establecidos para identificar contenido positivo, negativo o neutro pero el número específico otorgado es determinado por los investigadores.

Referente a los resultados de la comparación de palabras más frecuentemente repetidas para cada categoría, durante los experimentos quedó claro, mediante la comparación del caso de etiquetado manual y automático, que se tiene mayor coincidencia de palabras positivas. Con lo que se puede concluir que el clasificado de esta categoría es en su mayoría correcto. En el caso de los comentarios neutros y negativos, se observó una diferencia importante entre el etiquetado manual y el automático, una vez que se unificaron las clases neutro y positivo se pudieron encontrar más coincidencias entre los etiquetados, con esto se puede concluir que es más complicado diferenciar entre comentarios neutros y negativo que identificar la clase de comentarios positivo.

En cuanto al desempeño de la *Natural Language API* de Google, se concluye que la clasificación obtenida es comparable con la clasificación que haría un ser humano, resultando una herramienta útil para el análisis de sentimientos, vale la pena resaltar que los comentarios analizados variaban en su longitud, comprobándose que la herramienta de clasificación automática funciona para textos largos y cortos. Como se comentó anteriormente, la función de análisis de sentimientos es sólo una de las tareas que permite realizar la *API*, en el caso que abordó este trabajo se utilizaron comentarios en el idioma español, el cuál es uno de los idiomas compatibles con la funcionalidad de análisis de sentimientos, pero algunos otros métodos solo están habilitados para el idioma inglés. Entre los trabajos a futuro se propone la experimentación mediante los otros métodos disponibles en la *Natural Language API*, así como experimentación con otras herramientas de procesamiento de lenguaje natural de Google Cloud.

6. Conclusiones

El presente trabajo aporta una metodología de análisis de comentarios referentes a opiniones de turistas o potenciales turistas de un Área Natural Protegida en México. Siguiendo dicha metodología es posible identificar palabras frecuentes que usan los usuarios cuando se expresan de forma positiva o negativa de la región. Dichas palabras se pueden consultar en la Fig. 6 y en la Fig. 10.

Entre las palabras que destacan para la categoría positiva se encuentran: lugar, Tehuacán, hermoso, excelente, reserva, México, bonito, Puebla,

Naturaleza, conocer y país, lo que nos da una idea de los entes de los que se expresan de forma positiva.

Por otro lado para el caso de comentarios negativos, las palabras que se destacan son: más, cerrado, Tehuacán, tehuacancuicatlan, gobierno, región, zona, llegar, reserva y mundo, es posible que los usuarios estén haciendo referencia al estado de acceso a las zonas, ya que los últimos meses ha permanecido cerrada por motivos de la pandemia por la COVID-19.

Los resultados obtenidos muestran que el uso de estas herramientas para identificar los sentimientos de los usuarios de las ANP podrían ser provechosas para determinar mejoras en el aprovechamiento de dichas áreas.

También se compara la herramienta de procesamiento de lenguaje natural de Google para el análisis de sentimientos con un etiquetado manual y se concluye que su uso es útil para este tipo de aplicaciones.

Referencias

1. Cambria, E., Das, D., Bandyopadhyay, S., Feraco, A., et al.: A practical guide to sentiment analysis. Springer (2017)
2. Chaudhuri, A.: Visual and Text Sentiment Analysis through Hierarchical Deep Learning Networks. Springer (2019)
3. CONABIO: Áreas protegidas. https://www.biodiversidad.gob.mx/region/areasprot (Aug 2020), consultado: 2021-9-20
4. CONANP: Sistema de consulta fichas ANP. $https://simec.conanp.gob.mx/consulta_fichas.php, consultado : 2021 - 9 - 20$
5. CONANP: Marco estratégico de turismo sustentable en áreas naturales protegidas de méxico (METS). https://www.gob.mx/conanp/documentos/marco -estrategico-de-turismo- sustentable-en-areas-naturales-protegidas-de-mexico (Feb 2019), consultado: 2021-9-20
6. CONANP: Reserva de la biosfera tehuacán - cuicatlán. https://www.gob.mx/conanp/documentos/reserva-de-la-biosfera-tehuacan-cuicatlan-209465 (Sep 2019), consultado: 2021-9-20
7. CONANP: Turismo y naturaleza en áreas naturales protegidas. https://www.gob.mx/conanp/prensa/turismo-y- naturaleza-en-areas-naturales-protegidas-211386 (Aug 2019), consultado: 2021-9-20
8. CONANP: Programa nacional de áreas naturales protegidas 2020-2024. https://www.gob.mx/conanp/documentos/programa-nacional-de-areas-naturales -protegidas- 2020-2024 (Sep 2020), consultado: 2021-9-20
9. CONANP: Áreas naturales protegidas de méxico. http://sig.conanp.gob.mx/website/pagsig/ (Mar 2021), consultado: 2021-9-20
10. FLATICON: Free vector icons and stickers. https://www.flaticon.com/ (2010), consultado: 2021-11-18
11. Gobierno de México: Comisión nacional de áreas naturales protegidas. https://www.gob.mx/conanp, consultado: 2021-9-20
12. González Herrera, M.R., Miranda de Sá Teles, R., Bauer, R., Marques Gomes, C.: Tecnologias da informação e da comunicação e suas interfaces com o turismo: alguns estudos de caso brasil e méxico. Departamento de Ciencias Administrativas (2019)

13. GOOGLE: Cloud computing services. https://cloud.google.com/, consultado: 2021-9-20
14. GOOGLE: Google maps. https://www.google.com.mx/maps/, consultado: 2021-9-20
15. GOOGLE: How to use cloud natural language API. https://cloud.google.com/natural-language/docs/how-to, consultado: 2021-9-20
16. GOOGLE: Places API overview. https://developers.google.com/maps/, consultado: 2021-9-20
17. GOOGLE: YouTube data API. https://developers.google.com/youtube/v3/, consultado: 2021-9-20
18. Kamath, U., Liu, J., Whitaker, J.: Deep learning for NLP and speech recognition, vol. 84. Springer (2019)
19. Molinar, C.M.A., Espinoza, P.M., Llamas, I.O.: Evaluación de destinos turísticos mediante la tecnología de la ciencia de datos. Estudios y perspectivas en turismo **26**(2), 286–305 (2017)
20. Muthuraman, S., Al Haziazi, M.: Smart tourism destination-new exploration towards sustainable development in sultanate of oman. In: 2019 5th International Conference on Information Management (ICIM). pp. 332–335. IEEE (2019)
21. Pham, T.D., Vo, D., Li, F., Baker, K., Han, B., Lindsay, L., Pashna, M., Rowley, R.: Natural language processing for analysis of student online sentiment in a postgraduate program. Pacific Journal of Technology Enhanced Learning **2**(2), 15–30 (2020)
22. PYTHON: Welcome to python. https://www.python.org/, consultado: 2021-9-20
23. Sandoval Ortega, A.S., Alcalá De la O, B., Martínez Morales, J.: Marketing digital: Un análisis del consumidor en méxico. http://congreso.investiga.fca.unam.mx/docs/xxiii/docs/14.07.pdf (Oct 2018), consultado: 2021-9-20
24. SECTUR: Datatur3 - mapa del sitio. https://datatur.sectur.gob.mx, consultado: 2021-9-20
25. SECTUR: Boletín 81.- SECTUR y SEGITTUR trabajan coordinamente para desarrollar el primer destino inteligente mexicano. http://www.sectur.gob.mx/sala-de-prensa/2015/05/06/boletin-81-sectur-y -segittur-trabajan-coordinamente-para-desarrollar-el-primer-destino-inteligente-mexicano/ (May 2015), consultado: 2021-9-20
26. SECTUR: Cozumel se convertirá en el primer destino turístico inteligente de méxico. https://www.gob.mx/sectur/prensa/cozumel-se-convertira-en-el-primer-destino-turistico-inteligente-de-mexico (Dec 2015), consultado: 2021-9-20
27. SECTUR: El turismo es una de las actividades con más crecimiento en la economía nacional. https://www.gob.mx/sectur/prensa/el-turismo-es-una-de-las-actividades-con-mas-crecimiento-en-la-economia-nacional (Jul 2019), consultado: 2021-9-20
28. SECTUR: Atlas turístico de México. https://www.atlasturistico.sectur.gob.mx (2020), consultado: 2021-9-20
29. SETUR, SEGITTUR: Red de destinos turísticos inteligentes. https://www.destinosinteligentes.es/ (2019), consultado: 2021-9-20
30. UNAM: Infraestructura de datos espaciales abiertos. https://www.gits.igg.unam.mx/idea/descarga (2017), consultado: 2021-9-20
31. YouTube: YouTube. https://www.youtube.com/, consultado: 2021-9-21