TECNOLÓGICO NACIONAL DE MÉXICO,
CAMPUS CIUDAD GUZMÁN

PROGRAMA DE MAESTRÍA
EN CIENCIAS DE LA COMPUTACIÓN

DESARROLLO DE HERRAMIENTAS DE SOFTWARE PARA DETERMINAR EL PERFIL DEL TURISTA PARA GENERAR RECOMENDACIONES AUTOMÁTICAS

TESIS

que para obtener el grado de
MAESTRO EN CIENCIAS DE LA COMPUTACIÓN

presenta:

SAMUEL ARCE CARDENAS

Directores:

Dr. Daniel Fajardo Delgado

Dr. Miguel Ángel Álvarez Carmona

Zapotlán el Grande, Jalisco, México, Noviembre de 2021

Ciudad Guzmán, 25/noviembre/2021
DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN
Asunto: *Autorización de impresión de Tesis*

ING. SAMUEL ARCE CARDENAS
CANDIDATO AL GRADO DE MAESTRO EN CIENCIAS DE LA COMPUTACIÓN
PRESENTE

De acuerdo con los Lineamientos para la Operación de los Estudios de Posgrado en el Tecnológico Nacional de México y las disposiciones en este Instituto, habiendo cumplido con todas las indicaciones que la Comisión Revisora realizó con respecto a su Trabajo de Tesis titulado "Desarrollo de herramientas de software para determinar el perfil del turista para generar recomendaciones automáticas", la División de Estudios de Posgrado e Investigación de este Instituto, concede la Autorización para que proceda a la impresión del mismo.

Sin otro particular, quedo de Usted.

A T E N T A M E N T E
Excelencia en Educación Tecnológica
"Innova, Transforma y Crea para ser Grande"

EDUCACIÓN | TECNOLÓGICO NACIONAL DE MÉXICO
INSTITUTO TECNOLÓGICO DE CD. GUZMÁN
DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

DRA. MARIA GUADALUPE SÁNCHEZ CERVANTES
JEFA DE LA DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

ccp. Archivo
MGSS/megg

Resumen de la tesis de **Samuel Arce Cardenas**, presentada como requisito parcial para la obtención del grado de Maestro en Ciencias de la Computación. Zapotlán el Grande, Jalisco, México, Noviembre de 2021.

Desarrollo de herramientas de software para determinar el perfil del turista para generar recomendaciones automáticas.

Resumen aprobado por:

Dr. Daniel Fajardo Delgado	Dr. Miguel Ángel Álvarez Carmona
Director de Tesis	Codirector de Tesis

La presente tesis trata sobre la implementación de sistemas de recomendación turística diseñados para predecir las preferencias de los usuarios sobre un lugar o actividad turística en México. Se propusieron tres sistemas de recomendación: dos basados en filtrado colaborativo (usuario e ítems) y otro basado en filtrado demográfico. Se generó un corpus mediante la recopilación de 2,263 calificaciones de TripAdvisor.com sobre dieciocho lugares turísticos de México. Los resultados experimentales muestran que el sistema de recomendación con filtrado demográfico supera a los basados en el filtrado colaborativo, obteniendo un error absoluto medio de 0.67 y un error cuadrático medio de 1.2980. Estos resultados también muestran una mejora significativa sobre una línea de base de clase mayoritaria basada en un desequilibrio considerable.

Palabras clave: Sistema de recomendación turística, filtrado colaborativo, filtrado demosgráfico, turismo en México.

Abstract of the thesis presented by **Samuel Arce Cardenas**, in partial fulfillment of the requirements for the Master degree in Computer Science. Zapotlán el Grande, Jalisco, México, November 2021.

Development of software tools to determine the profile of the tourist to generate automatic recommendations.

Abstract approved by:

Dr. Daniel Fajardo Delgado	Dr. Miguel Ángel Álvarez Carmona
Director de Tesis	Codirector de Tesis

This thesis addresses the implementations of tourist recommendation systems designed to predict the user preferences about a place or tourist activity in Mexico. Three recommendation systems have been proposed: two based on collaborative filtering (user and items) and the other based on demographic issues. To this aim, a corpus has been built by collecting 2,263 ratings from TripAdvisor.com about eighteen tourist places in Mexico. Experimental results show that the demographic-based recommendation system outperforms those based on collaborative filtering, obtaining a mean absolute error of 0.67 and a mean square error of 1.2980. These results also show significant improvement over a majority class baseline based on a sizeable unbalanced corpus.

Keywords: Tourist recommender system, collaborative-based filtering, demographic-based filtering, tourism in Mexico.

Dedicatoria

A mis padres

Samuel y Guadalupe

Por su apoyo incondicional

Agradecimientos

A Dr. Daniel Fajardo Delgado y al Dr. Miguel Ángel Álvarez Carmona por su apoyo para realizar este trabajo de tesis. Su confianza y su capacidad han sido guía invaluable.

A los miembros del comité de tesis, la Dra. Rosa María Michel Nava y la M.C. Areli Pérez Aparicio, por su apoyo durante el proceso de revisión de este trabajo.

Al Consejo Nacional de Ciencia y Tecnología (CONACYT).

Al Tecnológico Nacional de México / Campus Cd. Guzmán.

Contenido

Página

Contenido (continuación)

Lista de figuras

Figura Página

Lista de tablas

Lista de algoritmos

Capítulo I

Introducción

El turismo en México tiene un impacto significativo en la economía por sus efectos multiplicadores en la generación de valor agregado y empleo (INEGI, 2019). Tan sólo en 2019, la industria turística aportó el 17.2 % del producto interno bruto (PIB) en México (El economista, 2019), obteniendo el sexto lugar en el ranking internacional de turismo (SECTUR, 2018). En términos de derrama económica, el turismo en México representa aproximadamente 22,500 millones de dólares al año (Bilal $et\ al.$, 2019). Recientemente, el impacto económico generado por la pandemia del coronavirus SARS-CoV-2 tiene repercusiones que pueden extenderse a medio plazo (Aguirre Quezada, 2020; EFE, 2020; Welle, 2020). A pesar de ello, las tecnologías digitales han permitido una reorientación de los modelos sociales, culturales y económicos relacionados con las propuestas turísticas que paliarían dicho impacto.

Una de las tecnologías donde se genera una gran cantidad de información útil al sector turista son las redes sociales. Las redes sociales son una herramienta que permite a los usuarios leer, escribir y expresar sus opiniones acerca de diferentes destinos y/o preferencias turísticas (Kandias $et\ al.$, 2017). Una forma de aprovechar la información que generan las redes sociales para el sector turismo es la creación de un perfil turista. En este contexto, el $perfil\ turista$ consiste de una amalgama de variables sociodemográficas y variables de comportamiento relacionados a actividades

turísticas (Li $et\ al.$, 2019). Con un perfil turista definido, es posible proveer al usuario de sugerencias y/o recomendaciones que sean de su interés para actividades turísticas.

Una de las tecnologías que optimizan el proceso de selección de lugares turísticos adecuados en función del perfil de usuario es el $sistema\ de\ recomendación$. Un sistema de recomendación turística busca predecir una "puntuación" o preferencia que los usuarios tienen con respecto a las opciones turísticas, con el objetivo de hacer coincidir las atracciones turísticas con las necesidades de los usuarios (Adomavicius y Tuzhilin, 2011). Estos sistemas buscan predecir un "puntaje" o preferencia que un usuario tiene respecto a una opción turística (Kzaz $et\ al.$, 2018). En el turismo, los sistemas de recomendación tienen como objetivo hacer coincidir las atracciones de los turistas con las necesidades del usuario (Adomavicius y Tuzhilin, 2011).

En el presente trabajo se propone diseñar e implementar un algoritmo de un corpus creado en redes sociales que pueda generar recomendaciones automáticas a turistas.

I.1. Objetivos

A continuación se presentan los objetivos de este trabajo de investigación.

I.1.1. Objetivo general

El objetivo general del presente trabajo fue diseñar e implementar un sistema que, con base en un perfil turista del usuario, genere recomendaciones automáticas de lugares turísticos en México.

I.1.2. Objetivos específicos

Los objetivos específicos fueron los siguientes:

- Buscar en el estado del arte un corpus de datos apropiado para el entrenamiento de un sistema de recomendaciones turísticas.
- Diseñar la representación del turista para utilizarla en sistemas de recomendación.
- Implementar los métodos de representación y clasificación de texto adecuados para el entrenamiento del sistema de recomendación.
- Implementar algoritmos basados en recomendar actividades turísticas según el perfil de usuarios en redes sociales.
- Desarrollar un prototipo del sistema de recomendación con los modelos desarrollados en los puntos anteriores.

I.2. Antecedentes

Actualmente, existe una gran variedad de trabajos en la literatura científica que tratan la aplicación de sistemas de recomendaciones turísticas.

Por ejemplo, Menk $et\ al.$ (2017) describen y evalúan un sistema capaz de generar recomendaciones turísticas fortuitas de ciudades de todo el mundo. Ellos utilizan el factor de curiosidad humana y lo combinan con características sociodemográficas obtenidas a través de la red social Facebook. Los autores evaluaron el sistema con la

participación de usuarios de Facebook obteniendo resultados satisfactorios en diversas métricas de desempeño (precisión, casualidad y novedad). Han y Lee (2015) proponen un enfoque que hace recomendaciones turísticas de forma adaptativa utilizando redes sociales geoetiquetadas. Primero examinan el impacto de las propiedades espaciales y temporales de un viaje considerando lugares populares. Después, calculan la importancia de los puntos de referencia para los viajeros en función de sus propiedades. Por último, generan grupos de recomendaciones históricas y un historial de trayectoria de viaje. Menk $et\ al.$ (2019) proponen un sistema de recomendaciones de viajes mediante el análisis de las redes sociales. Ellos muestran en su análisis que los lugares populares visitados por los viajeros cambian en función de las propiedades del viaje. Con base a este análisis, desarrollaron un software que recomienda puntos de referencia atractivos para los viajeros que usan las redes sociales. Además, es aplicable a lugares menos conocidos y refleja eventos locales y cambios estacionales.

Binucci $et\ al.$ (2017) diseñaron un analizador de contenido para un sistema de recomendación de viajes basado en contenido. Utilizan datos geográficos proporcionados por un conjunto de puntos de interés (PDI) para indicar qué tan relevante es un PDI para un conjunto de posibles temas de interés. Por otro lado, Vu $et\ al.$ (2019) obtienen las preferencias gastronómicas de los turistas en función de los sitios web de reseñas de restaurantes. Se utilizan técnicas de procesamiento de texto para analizar las preferencias de los turistas con respecto a las actividades gastronómicas (cocinas, platos, comidas y características del restaurante). Shen $et\ al.$ (2016) utilizan las redes sociales basadas en la ubicación para ofrecer a los turistas las recomendaciones de

lugares locales más relevantes y personalizadas. Al-Ghobari $et\ al.$ (2021) proponen un sistema de recomendación turística que integra las preferencias de los usuarios y su información geográfica para generar recomendaciones personalizadas y conscientes de la ubicación. Ellos usaron un filtrado colaborativo basado en elementos del algoritmo de aprendizaje k-Nearest Neighbor para este propósito. Su solución tenía como objetivo desarrollar una aplicación móvil que utiliza el servicio de Google para proporcionar sugerencias basadas en atracciones populares cercanas. Kuanr y Mohanty (2020) presentan un sistema de recomendación turística que almacena las opiniones de los usuarios locales sobre sus preferencias de comida y compra. Su sistema utiliza la información almacenada encontrando usuarios similares a cualquier usuario que consulta y brindándole recomendaciones de los sitios con buena comida y productos disponibles en esos sitios.

Fararni $et\ al.$ (2021) proponen una arquitectura híbrida y un marco conceptual basado en tecnologías de big data, inteligencia artificial e investigación operativa. Otros trabajos de investigación, como los revisados en Yochum $et\ al.$ (2020), también utilizan enfoques híbridos mediante el uso de datos abiertos vinculados (un concepto en el que los datos se comparten y construyen en base a la web semántica, datos vinculados y datos abiertos) en el ámbito turístico. Finalmente, Logesh $et\ al.$ (2019) proponen un enfoque híbrido para predecir las recomendaciones de puntos de interés persuasivos para una actividad y comportamiento personalizados inducidos.

I.3. Contribuciones

Las contribuciones principales de esta tesis son las siguientes:

- La creación de un corpus en español orientado al turismo que ayuda a la tarea de un sistema de recomendaciones turísticas. Este corpus se utilizó para una competencia en el taller Rest-Mex 2021[1].
- El análisis, diseño e implementación de tres tipos distintos de sistema de recomendaciones: colaborativo basado en usuarios, colaborativo basado en ítems y demográfico.
- Como un trabajo transversal al objetivo de la presente tesis, se participó en el MEX-A3T 2020 en las tareas de detección de noticias falsas y de identificación de agresividad en Twitter. Los resultados obtenidos se publicaron en (Arce Cardenas *et al.*, 2020). Este trabajo transversal se describe en el Apéndice A de este documento de tesis.
- La generación de conocimiento en el área de procesamiento de lenguaje natural y sistemas de recomendaciones.

I.4. Organización de la tesis

El resto del presente documento se organiza de la siguiente manera. El Capítulo II trata los conceptos y definiciones relacionadas a los sistemas de recomendación y su

[1] https://sites.google.com/cicese.edu.mx/rest-mex-2021/home

taxonomía. El Capítulo III describe las etapas de desarrollo de los modelos de sistemas de recomendación propuestos. El Capítulo IV presenta el diseño experimental y los resultados de desempeño de los modelos de sistemas de recomendación propuestos. Finalmente, el Capítulo V muestra las conclusiones y el trabajo futuro.

Capítulo II

Marco teórico

Los *sistemas de recomendación* son herramientas de software que filtran información con el fin de proporcionar sugerencias útiles para un usuario (Ricci *et al.*, 2011). Estas sugerencias generalmente hacen referencia a ítems (o elementos) que se pueden clasificar según su complejidad, valor o utilidad. El sistema de recomendación puede utilizar diferentes atributos y características del ítem según su tecnología. Por ejemplo, en un sistema de recomendación de películas, tanto el género como el director y los actores pueden usarse para describir la película y comprender cómo la utilidad del ítem depende de sus características.

Este tipo de sistemas buscan estimar datos perdidos de los ítems que el usuario aún no ha contemplado (y por lo tanto desconocemos el nivel de interés del usuario sobre ellos) (González y Jacques, 2017; Amatriain, 2014). En general, dentro de los sistemas de recomendación, no sólo se escoge un único elemento, si no que se crea un *ranking* de elementos basándose en el nivel de interés estimado y se seleccionan los mejores (Adomavicius y Tuzhilin, 2005).

Existen diferentes técnicas para la generación de modelos de sistemas de recomendación, entre las que destacan las basadas en contenido, de filtración colaborativa, las de filtrado demográfico y las híbridas. A continuación se describen cada una de ellas.

II.1. Recomendaciones basadas en contenido

En los sistemas de recomendaciones basados en contenido, un usuario recibe recomendaciones de elementos similares a los que el usuario prefirió en el pasado. La toma de decisiones en este tipo de sistemas considera diferentes factores de contenido a partir de las preferencias de los turistas (Hamid $et\ al.$, 2021).

En este enfoque, las recomendaciones se construyen a partir de la recopilación de información acerca del comportamiento del usuario (Burke, 2002). Dicha información puede obtenerse de forma implícita o explícita. Cuando la información se obtiene de forma implícita, generalmente se utilizan técnicas de aprendizaje máquina capaces de identificar automáticamente el perfil turístico del usuario. Por otro lado, cuando la información se obtiene de forma explícita, la información del usuario generalmente se recopila directamente de él a través de formularios, encuestas, opiniones, entre otros.

De acuerdo con Ricci $et\ al.$ (2011), cuando cada ítem se describe con el mismo conjunto de atributos y hay un conjunto conocido de valores para esos atributos, entonces el ítem tiene una representación estructurada. Por otro lado, existen descripciones textuales de los ítems, como las tomadas de sitios web, correos electrónicos, artículos de noticias o descripciones de productos. Algunos atributos tienen valores definidos explícitamente. La representación textual crea mucha complejidad debido a la ambigüedad del lenguaje natural. El problema es que la configuración basada en palabras clave se basa principalmente en la similitud entre cadenas, por lo que los usuarios no pueden capturar la semántica de interés. Según Ricci $et\ al.$ (2011), hay

dos problemas con la similitud de cadenas: ambigüedad (múltiples significados de una palabra) y sinónimos (varias palabras con el mismo significado).

Las ventajas del sistema de recomendaciones basadas en contenido son las siguientes:

- Las recomendaciones se generan a partir del usuario, sin requerir de opiniones subjetivas de otros usuarios. A esta característica se le conoce como “sin dispersión”.
- El sistema puede generar descripciones recomendadas basadas en el historial del usuario.
- El modelo de información está incluido en las propiedades de cada documento que se evalúa.

A continuación se listan algunas desventajas:

- Requiere un modelo configurado por el usuario, el cuál generalmente es complejo de construir y mantener.
- Los usuarios están limitados a artículos recomendados similares a los recomendados por él.
- Se presentan dificultades cuando el contenido es difícil de analizar (audio, gráficos, imágenes, vídeo).
- Los usuarios deben contener los elementos suficientes para que el sistema identifique efectivamente sus preferencias.

- Se extraen características específicas de cada elemento para evaluar la similitud.
- Se presenta un “efecto billetera”, donde se pueden rechazar contenidos muy similares a los ya evaluados.
- Típicamente, el sistema tiene dificultades para adaptarse a los cambios en los perfiles de usuario hasta que se haya recopilado una cantidad suficiente de exámenes actualizados.

Finalmente, el filtrado basado en contenido no es un enfoque adecuado cuando hay una ausencia de datos de usuarios previos para tomar decisiones. Bajo esta perspectiva, el filtrado colaborativo puede ofrecer datos iniciales basados en las similitudes de los usuarios (Xiong $et\ al.$, 2017).

II.2. Filtrado colaborativo

En el filtrado colaborativo, un usuario recibe recomendaciones de ítems estableciendo relaciones con personas con gustos y preferencias similares preferidas en el pasado. La mayoría de los métodos de filtrado colaborativo solo acceden al ID de usuario y del ítem, sin más información sobre ellos. Existe una matriz de elementos de usuario donde cada entrada en la matriz puede ser un valor desconocido o una calificación de usuario para el ítem, esta última generalmente asignada en una escala particular.

Zafarani $et\ al.$ (2014) presentan dos enfoques para las técnicas de filtrado colaborativo: basado en memoria y basado en modelo. Los algoritmos de filtrado

colaborativo basados en memoria suponen uno o ambos de los siguientes hechos:

1. Usuarios que comparten un historial de ratings similares es probable que califiquen de manera similar a otros items en el futuro.

2. Los ítems que comparten ratings similares en el pasado, son propensos a recibir ratings similares de los futuros usuarios.

En el primer hecho se refiere a la técnica basada en memoria de un filtrado colaborativo basada en el usuario y la segunda se refiere a un filtrado colaborativo basado en el ítem. En los dos hechos los usuarios e ítems trabajan juntos para eliminar contenido irrelevante. Una de las medidas de similitud que es utiliza para determinar la similitud entre usuarios o ítems es la de coseno.

Para los algoritmos de filtrado colaborativo basados en modelos, se da por hecho que el modelo subyacente controla cómo se evalúa a los usuarios. Por lo tanto, el desafío es comprender el modelo y usarlo para predecir la evaluación. Las técnicas de aprendizaje que comúnmente se utilizan son: Bayesian Clustering (Breese $et\ al.$, 2013), Hofmann Latent Semantics (Hofmann, 2003), Latent Dirichlet Allocation (Blei $et\ al.$, 2003), entropía máxima (Zitnick y Kanade, 2004), máquinas de Boltzmann (Salakhutdinov $et\ al.$, 2007), y máquinas de vector soporte (Grčar $et\ al.$, 2006).

Las ventajas del filtrado colaborativo son las siguientes:

- No necesita de un modelo detallado de las preferencias del usuario, con una matriz, $user\text{-}ítem$ es suficiente.

- Permite sugerencias de contenido que es difícil de analizar.
- Recomendaciones de artículos según las preferencias del usuario.
- Aplicable a todo tipo de artículos y productos, incluidos documentos, música, películas y libros.
- Puede introducir nuevas funciones vinculadas a la experiencia de usuario anterior.
- Similar a la popularidad global, pero personalizado para los usuarios (en comparación con las “puntuaciones”, estos son otros usuarios).

A continuación se listan algunas desventajas:

- Problema de dispersión: si el número de usuarios es pequeño en comparación con la cantidad de información en el sistema, existe el riesgo de que el alcance de la evaluación esté muy fragmentado. Reducir la colección de artículos sugeridos.
- Generalmente se presenta un “problema de arranque frío” este problema sucede cuando se agrega un nuevo usuario o un nuevo ítem, esto quiere decir que es más difícil recomendar a un nuevo usuario o recomendar un ítem que acaba de ser agregado.
- Es difícil identificar las recomendaciones adecuadas porque algunos usuarios tienen perfiles que pertenecen a una clase de usuario existente.
- Problema de sinónimos: se produce debido a una falta de interpretación semántica de todas las formas. Una vez que se hace una sugerencia, los artículos similares

no se tratarán de esta manera.

II.3. Filtrado demográfico

Esta técnica clasifica a los usuarios según su perfil y hacen recomendaciones según sus clases demográficas. Las recomendaciones demográficas son similares a las recomendaciones basadas en contenido, excepto que la similitud se calcula utilizando información demográfica en lugar de los ítems.

La figura I sugiere utilizar los datos demográficos de los usuarios almacenados en sus perfiles (es decir, edad, sexo, ubicación, etc.), asume que los usuarios con atributos demográficos similares calificarán los artículos de manera similar. Este recomendador obtiene un grupo de usuarios con atributos demográficos similares que forman un modelo a partir del cual se generan los nuevos ítems recomendados.

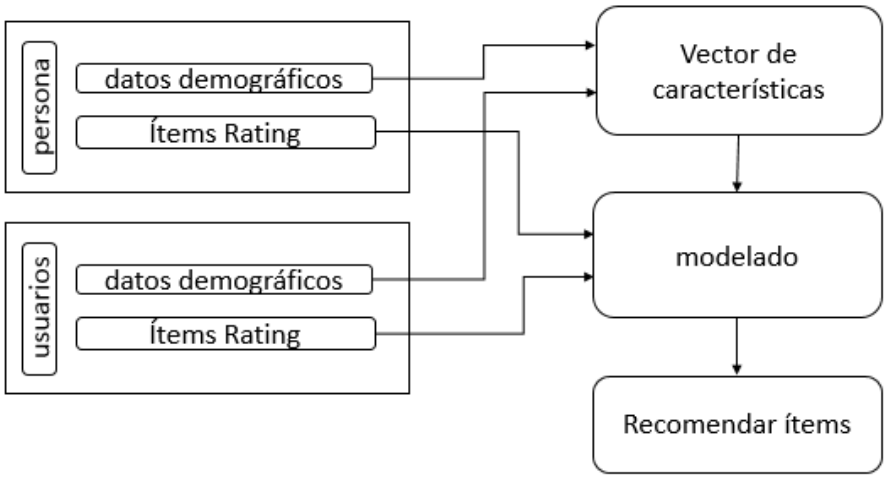

Figura I. Diagrama de flujo filtrado demográfico

II.4. Algoritmos híbridos

Estos algoritmos se basan en una combinación de las técnicas anteriores y otras técnicas de recomendaciones. La idea es aprovechar un método y llenar los vacíos en el otro. Por ejemplo, los métodos de filtrado colaborativo presentan problemas cuando se insertan nuevos elementos. Sin embargo, el filtrado basado en contenido no se limita a este tipo de situación porque se basa en una descripción de objetos de fácil acceso. Hay varias formas de combinar estas técnicas. Burke (2002) hace una interesante recopilación de ellas y las resume de la siguiente manera:

- **Ponderado**. En un sistema de recomendación ponderado por combinación, la puntuación de recomendación se calcula a partir de los resultados de todas las técnicas recomendadas disponibles en el sistema. Por ejemplo, podría ser una combinación lineal de puntuaciones de cada una de las técnicas recomendadas implementadas.

- **Conmutación**. En este caso, el sistema cambiará el método recomendado según la situación actual. Por ejemplo, una técnica híbrida que combina el filtrado colaborativo y el basado en contenido, este último, respectivamente, se aplica primero. Si el sistema basado en contenido no proporciona suficiente confianza para las recomendaciones, se aplica un sistema de filtrado colaborativo. Estos sistemas combinados aumentan la complejidad del proceso de recomendación ya que deben definirse los criterios de conmutación, lo que introduce otro nivel de parametrización. La ventaja, sin embargo, es que el sistema puede ser sensible a

las fortalezas y debilidades de las tecnologías que lo componen.

- **Mixta**. Estos sistemas ofrecen recomendaciones para varias técnicas a la vez.
- **Combinación de atributos**. Una forma de combinar el filtrado colaborativo con el filtrado basado en contenido es tratar la información colaborativa como una característica de datos adicional simple asociada con cada ejemplo y utilizar técnicas basadas en contenido en el conjunto de datos desarrollado. Esta técnica permite revisar los datos de forma conjunta sin depender completamente de los datos, haciendo que el sistema sea menos sensible al número de usuarios que puntúan el artículo. Por otro lado, este método permite que el sistema tenga información sobre las similitudes intrínsecas de los elementos que de otro modo quedarían enmascarados por el sistema colaborativo.
- **Cascada**. En esta técnica, se aplica primero la técnica recomendada para generar una lista de ítems candidatos, y luego se aplica la segunda técnica para mejorar la propuesta generada previamente.
- **Incremento de atributos**. En este caso, la salida de una técnica se utiliza como característica de entrada de la otra técnica. Se utiliza esta técnica para generar calificaciones o calificaciones de ítems e incorporar esa información en el siguiente proceso de recomendación. Esta técnica es interesante porque permite mejorar el rendimiento del sistema sin modificar el sistema. Los intermediarios agregaron funcionalidad adicional que utilizaron otras técnicas para consolidar la información.

En particular destacan las dos primeras, respectivamente. Los sistemas de recomendaciones *basados en conocimiento* aprenden a recomendar elementos similares a los que le han gustado al usuario en el pasado. La similitud de los elementos se calcula en función de las características asociadas a los elementos comparados. Después, se identifican las características comunes de los elementos que han recibido una calificación favorable de un usuario, y luego recomienda nuevos elementos que compartan esas características (Billsus y Pazzani, 2000). Por otro lado, los sistemas de recomendaciones *basados en filtración colaborativa*, consisten en el proceso de filtrar o evaluar elementos utilizando las opiniones de otras personas (Schafer *et al.*, 2007). Estas opiniones se pueden obtener explícitamente de los usuarios a través de la respuesta del formulario, o mediante el uso de algunas medidas implícitas, como registros de compras anteriores (Ricci *et al.*, 2011).

Capítulo III

Metodología

En el presente trabajo se proponen tres sistemas de recomendación: filtrado colaborativo basado en los usuarios, filtrado colaborativo basado en los ítems y filtrado demográfico. En este capítulo se describen las fases de desarrollo de estos sistemas.

III.1. Construcción del corpus

Como un primer paso en el desarrollo de los sistemas de recomendación, se construyó un corpus mediante la recopilación de reseñas turísticas en español y valoraciones globales de dieciocho lugares o atracciones turísticas en el estado de Nayarit, México. Los dieciocho lugares turísticos se seleccionaron teniendo en cuenta ocho tipos de destinos turísticos: sol y playa, cultural, aventura, religioso, natural, gastronómico, ecoturismo y compras. Tanto para la selección como la definición de la tipología de los lugares, se recibió el apoyo del Dr. Juan Pablo Ramírez Silva, investigador de la Universidad Autónoma de Nayarit. La tabla I muestra la tipología de turismo utilizada para estos lugares.

Los datos se obtuvieron de TripAdvisor.com, un sitio Web cuyo contenido plasma comentarios de las experiencias de viaje de los usuarios. Para llevar a cabo la recopilación de la información se utilizó un rastreador Web (o $Web\ crawler$) a través de las siguientes

Tabla I. Tipología de lugar turístico

Lugares turísticos	**Tipología de lugar turístico**
Bahía de Matanchen	sol y playa
Playa Los Muertos	sol y playa
Bucerias Art Walk	cultural, compras
Centro Histórico de Tepic	cultural, religioso
Galerías Vallarta	compras
Isla de Coral	sol y playa, ecoturismo
Islas Marietas	sol y playa, aventura, ecoturismo
Manantial La Tovara	sol y playa, aventura, ecoturismo
Mercado del Pueblo Sayulita	gastronómico
Mexcaltitan	natural
Playa Destiladeras	sol y playa
Playa El Anclote	sol y playa
Playa Los Ayala	sol y playa
Splash Water Park	aventura, compras
The Jazz Foundation	cultural
Isla Isabel	sol y playa, ecoturismo
Cerro de la Contaduría	cultural, aventura
Santuario de Cocodrilos El Cora	ecoturismo

dos herramientas de software: Selenium WebDriver y Python Selenium (Salunke, 2014). Se recolectaron un total de 2,263 opiniones y $ratings$ realizadas por 2,033 usuarios desde Mayo de 2012 hasta Enero de 2021. Cada una de estos $ratings$ consiste de una escala $Likert$ de cinco puntos: 1 (pésimo), 2 (malo), 3 (regular), 4 (muy bueno) y 5 (excelente). La opiniones recolectadas son el resultado de todos los comentarios encontrados en los 18 lugares etiquetados. La tabla II muestra la distribución de las instancias del corpus según su $rating$.

Además de las reseñas y valoraciones, también se obtuvo información sobre cada uno de los 2,033 usuarios a través del rastreador Web. Sin embargo, fue necesario un tratamiento manual de los datos para obtener adicionalmente el género de los usuarios

Tabla II. Distribución de las instancias del corpus según su $rating$.

rating	**Numero de instancias**
1	65
2	77
3	239
4	653
5	1229

y una breve opinión sobre los lugares valorados.

Cabe resaltar que no hay campos vacíos en los registros del corpus, de manera que si un usuario carecía de información (lugar de origen, tipo de viaje o género) se omitía el registro de dicho usuario y sus valoraciones. Por último, el nombre del usuario se reemplazó por un identificador preservando la privacidad de las opiniones. La tabla III muestra la información que se recopiló para cada usuario, conformando un registro de ocho campos.

Además, se recogió un historial de opiniones de algunos de los 2,033 usuarios, el cuál consiste de comentarios y observaciones que cada uno de estos usuarios hizo sobre los lugares turísticos que visitó (no necesariamente los que figuran en la tabla I). La tabla IV muestra los campos del historial de opiniones.

Finalmente, el conjunto de instancias del corpus se dividió en los siguientes dos grupos: una muestra de entrenamiento formada por 1,582 valoraciones seleccionadas aleatoriamente, y una muestra de prueba que contiene 681 instancias seleccionadas aleatoriamente para medir el rendimiento. La división de las instancias del corpus se realizó con validación estratificada $K\text{-}fold$, con base en la distribución de la tabla II. Por tanto, se aseguró el 70 % de cada valoración para la muestra de formación y el 30 %

Tabla III. Información del usuario en el corpus.

Campo	**Descripción**	**Tipo**
ID	El identificador del usuario por cada recomendación.	texto
Género	El género del turista	{hombre, mujer}
Lugar	El lugar turístico que el usuario visitó	texto
Locación	Lugar de origen del turista (centro, noreste, noroeste, este, y sureste hablan sobre regiones de México).	texto
Fecha	Fecha de la cuando la opinión fue hecha.	fecha
Tipo	Tipo de viaje que usuario hizo.	{familia, amigos, solitario, pareja, negocios}
rating	El *rating* representa el nivel de satisfacción que el turista tendría cuando viaja a un lugar	{1, 2, 3, 4, 5}
Comentario	El comentario que el turista otorga.	texto

Tabla IV. Historial de opiniones de cada usuario.

Campo	**Descripción**	**Tipo**
Comentario	El comentario que otorgó el usuario (desconocido = comentario en blanco)	texto
rating	El nivel de satisfacción que tuvo el usuario con respecto a un lugar específico.	{1, 2, 3, 4, 5}
Lugar	El lugar que visitó un usuario (este lugar puede ser de cualquier parte del mundo, no necesariamente de México)	texto
Locación	El lugar de origen del usuario (las regiones central, noreste, noroeste, oeste y sureste se refieren a las regiones de México).	texto
rating promeido	La calificación general que tiene un lugar en el sitio de TripAdvisor.com	[1..5]

para la muestra de prueba. A continuación, se describen los modelos utilizados para generar las recomendaciones de los lugares turísticos.

III.2. Generación de los sistemas de recomendación colaborativos

Se generaron dos modelos de sistemas de recomendación colaborativos: uno basado en usuarios y el otro en ítems. Mientras que el enfoque basado en usuarios encuentra a los usuarios que comparten los mismos patrones de valoración con el usuario objetivo, el enfoque basado en ítems examina el conjunto de artículos que el usuario objetivo ha valorado y calcula su similitud (ver Sección II.2). Los modelos de recomendación bajo estos dos enfoques, generalmente se construyen utilizando el conocido algoritmo k-Nearest Neighbors (KNN) o alguna de sus variantes.

Los algoritmos de tipo KNN permiten ofrecer recomendaciones mediante la agregación de las valoraciones de los vecinos k más cercanos. En particular, el presente trabajo utiliza el algoritmo de KNN with means de (Hedlund y Nilsson Tengstrand, 2020), que tiene en cuenta la calificación media de cada usuario así como la media de k vecinos. Dado que el número de artículos es bajo en comparación con los usuarios, se utilizaron diferentes valores de k; en particular se utilizaron $k = 10, 20, 25, 30, 35$. Estos valores se eligieron siguiendo el trabajo de Ghazanfar y Prugel-Bennett (2010), donde se evaluó el valor óptimo de k para un sistema de recomendación. Por otro lado, para el enfoque basado en ítems, los valores de k fueron $K = 1, 3, 5, 7, 9$.

Los algoritmos KNN para los enfoques basados en usuarios y en ítems se implementaron utilizando la biblioteca Surprise, que es un Scikit de Python para sistemas de recomendación. Esta biblioteca ofrece una gama de algoritmos de sistemas de recomendación, incluyendo variaciones del KNN y diferentes índices de similitud. La biblioteca Surprise también se utilizó para calcular los cuatro pasos siguientes (1) construir la matriz de valoración usuario-ítem, (2) calcular la matriz de similitud, (3) calcular las predicciones de las calificaciones e (4) identificar las recomendaciones.

III.2.1. Construcción de la matriz usuario-ítem

La matriz de valoración usuario-ítem se construye mediante las valoraciones hechas por los usuarios a los ítems (en este caso, los lugares turísticos). En esta matríz se consideran las similitudes entre las valoraciones de los usuarios para predecir las valoraciones de un usuario objetivo sobre determinados ítems. La tabla V muestra un ejemplo de la matriz de valoración usuario-ítem. En esta tabla, las columnas corresponden a los lugares turísticos y las filas a los usuarios. La intersección entre ellas es la valoración que un usuario da a un lugar turístico concreto.

Tabla V. Ejemplo de la matriz usuario-ítem.

	Islas Marietas	Manantial La Tovara	Sayulita	...	Mexcaltitan
user_1		5		...	
user_2	3			...	4
user_3		3		...	5
⋮	⋮	⋮	⋮	⋮	⋮
user_n	4		5	...	

III.2.2. Creación de la matriz de similitud

La matriz de similitud está formada por pesos que representan la relación entre dos elementos (usuarios e ítems). Cuanto mayor sea el valor del peso, más firme será la relación entre ellos. En este trabajo, se calculó la similitud del coseno entre todos los pares de elementos (usuarios o artículos) para generar los valores de pesos. Sea U_{ij} el conjunto de todos los usuarios que han valorado ambos elementos i y j en el sistema de recomendación, y sea I_{uv} el conjunto de elementos valorados por dos usuarios u y v. La valoración del usuario u para el artículo i se denomina r_{ui}.

La ecuación (1) expresa la similitud del coseno entre los usuarios u y v, mientras que la ecuación (2) describe la similitud entre los artículos i y j. La tabla VI muestra un ejemplo de matriz de similitud para los usuarios.

$$sim(u,v) = \frac{\sum_{i \in I_{uv}} r_{ui} \cdot r_{vi}}{\sqrt{\sum_{i \in I_{uv}} r_{ui}^2} \cdot \sqrt{\sum_{i \in I_{uv}} r_{vi}^2}} \tag{1}$$

$$sim(i,j) = \frac{\sum_{u \in U_{ij}} r_{ui} \cdot r_{uj}}{\sqrt{\sum_{u \in U_{ij}} r_{ui}^2} \cdot \sqrt{\sum_{u \in U_{ij}} r_{uj}^2}} \tag{2}$$

Tabla VI. Ejemplo de matriz de similitud para el enfoque basado en usuarios.

	User_1	User_2	User_3	...	User_n
User_1	-	0	0.61	...	0
User_2	0	-	0.45	...	0.4
User_3	0.61	0.45	-	...	0
⋮	⋮	⋮	⋮	⋮	⋮
User_n	0	0.4	0	...	-

III.2.3. Cálculo de las predicciones de los *ratings*

La predicción de los $ratings$ se calcula teniendo en cuenta el $rating$ promedio de cada usuario. Sea μ_u el $rating$ promedio de cada usuario u (o μ_i si la predicción se calcula utilizando el enfoque basado en ítems). La predicción del $rating$ $\hat{r}_{ui}$ para el usuario u sobre el ítem i se expresa en la ecuación (3) bajo el enfoque basado en el usuario y en la ecuación (4) bajo el enfoque basado en el artículo. En estas ecuaciones, $N_i^k(u)$ denota el conjunto de k vecinos de u que han valorado el ítem i.

$$\hat{r}_{ui} = \mu_u + \frac{\sum_{v \in N_i^k(u)} sim(u,v) \cdot (r_{vi} - \mu_v)}{\sum_{v \in N_i^k(u)} sim(u,v)} \tag{3}$$

$$\hat{r}_{ui} = \mu_i + \frac{\sum_{j \in N_u^k(i)} sim(i,j) \cdot (r_{uj} - \mu_j)}{\sum_{j \in N_u^k(i)} sim(i,j)} \tag{4}$$

El algoritmo 1 describe el proceso de filtrado colaborativo basado en usuarios. La líneas 1 y 2 del algoritmo, consideran cada elemento $item_i$ e $item_j$ de cada usuario $user_u$ para calcular las similitudes los ítems (ver línea 3). Para cada $item_i$, se agrega la preferencia del usuario $user_u$ por el $item_j$ con una ponderación (ver línea 5). Finalmente, el algoritmo regresa los ítems con mayor ponderación (ver línea 7) y que se recomendarán al usuario objetivo. La figura II muestra el diagrama de flujo de este algoritmo, donde se señalan los pasos que sigue el sistema de recomendación de filtrado colaborativo basado en usuarios.

El algoritmo 2 describe el filtrado colaborativo basado en ítems. Este algoritmo es

Algoritmo 1: Filtrado colaborativo basado en usuarios

```
1 para todos los item_i que el user_u no tiene preferencia hacer
2     para todos los item_j que el user_u tiene preferencia hacer
3         calcular similitudes entre item_i e item_j
4     fin
5     agregar la preferencia del user_u por el item_j ponderada por una ejecución
6 fin
7 devolver ítems principales, calificados por la ponderación promedio
```

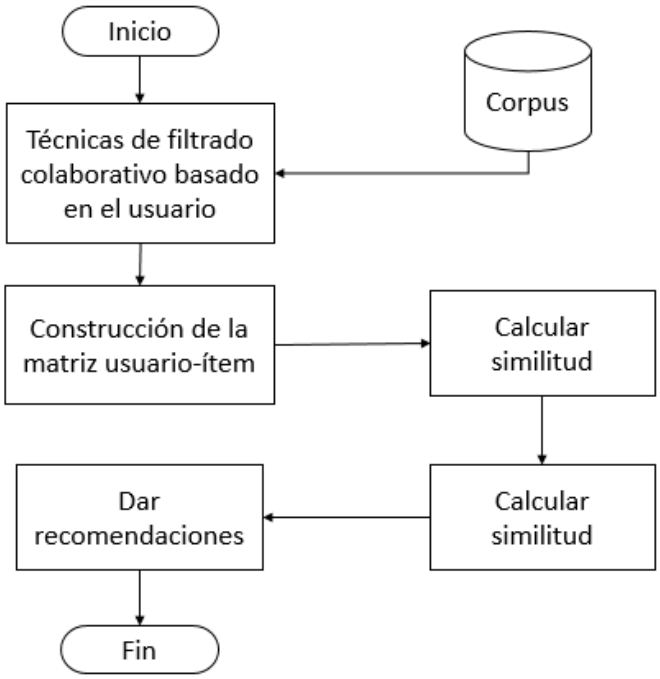

Figura II. Diagrama de flujo del filtrado colaborativo basado en el usuario

similar al algoritmo 1 con la diferencia en las líneas 2, 3 y 5. Las líneas 2 y 3 del algoritmo 2, consideran cada usuario $user_u$ y $user_v$ de cada $item_i$ para calcular las similitudes entre los usuarios. La línea 5 agrega la preferencia del usuario $user_v$ por el $item_i$. La figura III muestra el diagrama de flujo de este algoritmo 2, donde se señalan

los pasos que sigue el sistema de recomendación de filtrado colaborativo basado en ítems.

Algoritmo 2: Filtrado colaborativo basado en ítems

```
para todos los item_i que el user_u no tiene preferencia hacer
    para todos los user_v que tiene preferencia por el item_i hacer
        calcular similitudes entre user_u e user_v
    fin
    agregar la preferencia del user_v por el item_i ponderada por una ejecución
fin
devolver ítems principales, calificados por la ponderación promedio
```

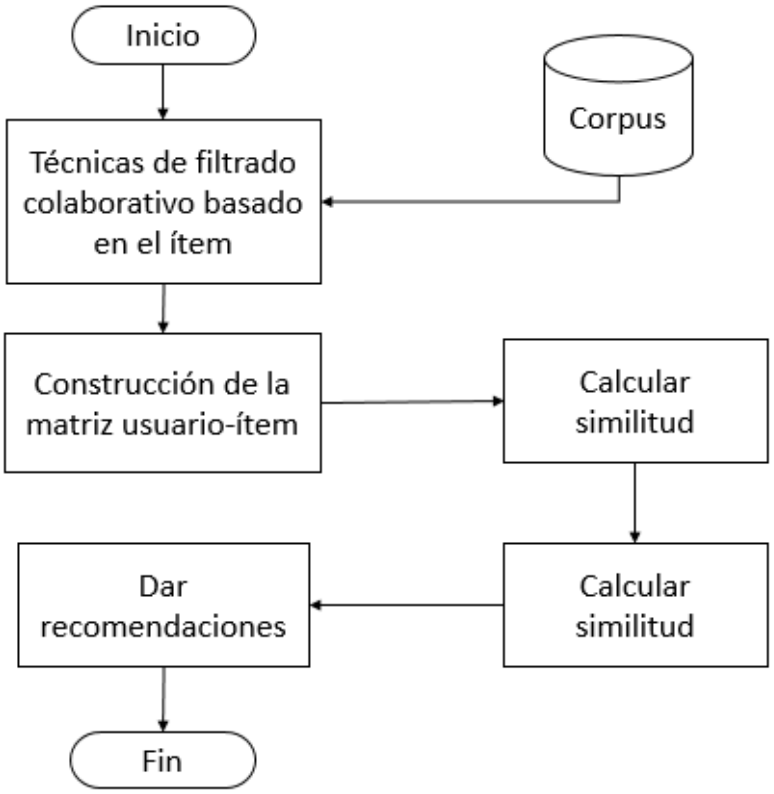

Figura III. Diagrama de flujo del filtrado colaborativo basado en el ítem

III.2.4. Filtrado demográfico

El filtrado demográfico clasifica a los usuarios o ítems en función de sus atributos y realiza una recomendación con base en la información demográfica (ver Sección II.3. En el presente trabajo, se generó un modelo de recomendación utilizando el filtrado demográfico con los siguientes tres atributos del usuario: género, ubicación y el tipo de viaje que realizó el turista. La tabla III muestra los campos que constituyen el registro de cada usuario. Por otro lado, la tabla VII muestra un ejemplo del conjunto de atributos que caracterizan la información demográfica de los usuarios. En cuanto a los registros de los lugares turísticos, cada uno de ellos consta de una representación binaria de rasgos de la tipología de destinos turísticos (ver tabla I). En esta representación binaria, los 0s y 1s indican si el lugar turístico se ajusta o no a uno o varios tipos de destinos turísticos, respectivamente. Para construir los modelos de recomendación, se utilizan los siguientes algoritmos de aprendizaje automático implementados en Python Scikit-Learn: el KNN para $k = 10, 20, 25, 30, 35$, los bosques aleatorios (RF, por sus siglas en inglés) y las redes neuronales (NN, por sus siglas en inglés).

Tabla VII. Información demográfica de los usuarios que calificaron un lugar turístico.

ID	Genero	Lugar	Tipo de viaje
User_1	Hombre	Argentina	Negocios
User_2	Mujer	West region of Mexico	Familia
User_3	Mujer	Central region of Mexico	Solo
⋮	⋮	⋮	⋮
User_n	Hombre	USA	Amigos

La figura IV muestra el diagrama de flujo del sistema de recomendaciones demográfico con filtrado demográfico.

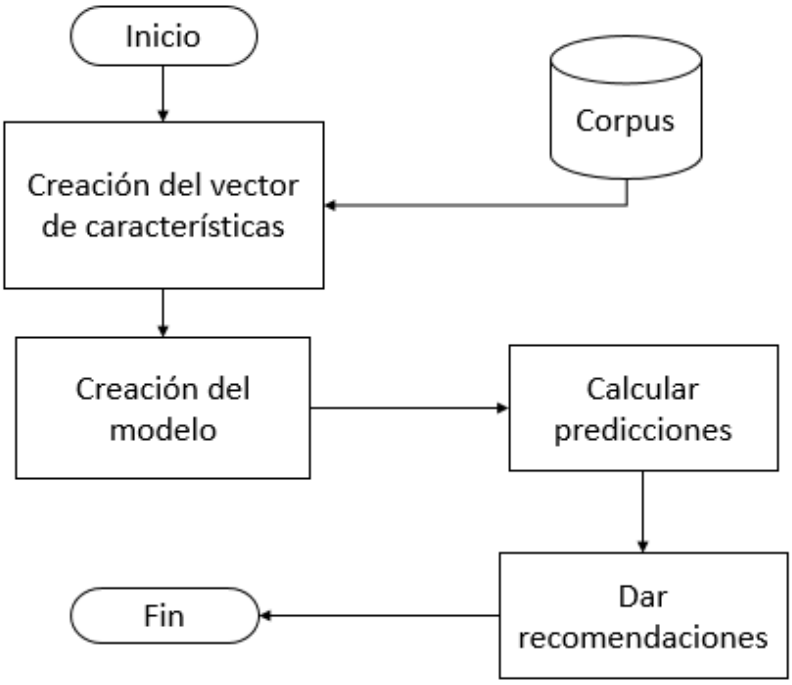

Figura IV. Diagrama de flujo filtrado demográfico

Capítulo IV

Resultados experimentales

En este capítulo se presentan los resultados de una evaluación experimental acerca del desempeño de los sistemas de recomendación propuestos. Para llevar a cabo esta evaluación, se utilizó una muestra de 681 registros del corpus que se utiliza como conjunto de pruebas para los sistemas de recomendación (ver sección III.1). Tanto el entrenamiento de los modelos de los sistemas de recomendación como la evaluación con el conjunto de pruebas, se llevaron a cabo en la plataforma de Google Colab, un servicio gratuito en la nube que incorpora elementos de código abierto y de texto en un Jupyter Notebook[1]. Se utilizaron un conjunto de métricas de desempeño que permiten comparar objetivamente el desempeño de los modelos.

IV.1. Métricas de evaluación

Las métricas de desempeño que se utilizan en este trabajo son el error absoluto medio (o MAE, $mean\ absolute\ error$), el error cuadrático medio (o MSE, $mean\ squared\ error$) y la raíz cuadrada del error cuadrático medio (o RMSE, $root\ mean\ squared\ error$).

Sea $r_{i\alpha}$ el rating del usuario u_i acerca del ítem v_α; y sea $r^o_{i\alpha}$ una predicción del rating con respecto a un ítem del conjunto de pruebas R_{TEST}. Los resultados de MAE, MSR

[1] https://colab.research.google.com

y RMSE se pueden obtener mediante las siguientes ecuaciones:

$$\text{MAE} = \frac{1}{|R_{TEST}|} \sum_{r_{i\alpha} \in N} |r_{i\alpha} - r^o_{i\alpha}| \tag{5}$$

$$\text{MSE} = \frac{1}{|R_{TEST}|} \sum_{r_{i\alpha} \in N} (r_{i\alpha} - r^o_{i\alpha})^2 \tag{6}$$

$$\text{RMSE} = \sqrt{MSE} \tag{7}$$

MAE es una medida de diferencia entre el $rating$ predicho y el real (o de prueba). El MSE calcula el promedio de los errores al cuadrado, lo que permite evaluar la calidad del modelo para realizar predicciones en cuanto a variación y sesgo. Finalmente, el RMSE es la raíz cuadrada del MSE, de forma que permite evaluar el impacto de errores con una magnitud mayor. Al comparar los ratings de prueba con los ratings predichos por cada modelo de recomendación, se tiene que cuanto más pequeños son los valores de MAE, MSE y RMSE, más precisa es la predicción del modelo de recomendación.

Con el objetivo de emplear otras métricas que permitan evaluar otros aspectos de los modelos de recomendación, se transformaron los valores de reales a enteros a través del redondeo. Así, se convirtió el problema de regresión original a un problema de clasificación considerando cinco clases: 1 (pésimo), 2 (malo), 3 (regular), 4 (muy bueno) y 5 (excelente).

Cada valor de predicción en comparación con cada registro del conjunto de pruebas,

puede arrojar cualquiera de las siguientes cuatro posibilidades: verdadero positivo (TP), verdadero negativo (TN), falso positivo (FP) y falso negativo (FN). Las métricas de desempeño que se utilizan en este trabajo son tres: la precisión, la sensibilidad (o $recall$) y el puntaje F1. La precisión consiste en la fracción de predicciones en las que el modelo acierta. Para el cálculo de la precisión se utiliza la ecuación (8). Por otro lado, la sensibilidad consiste de la fracción del total de registros relevantes que fueron obtenidos, y se calcula con la ecuación (9). La métrica exactitud mide el porcentaje de casos que el modelo ha acertado. Por otra parte, la exactitud del modelo se representa calculando el porcentaje de predicciones correctas. Esta métrica se expresa en la ecuación 10. Finalmente, el puntaje F1 representa la media armónica de precisión y sensibilidad, ver ecuación (11).

$$precisión = \frac{TP}{TP + FP} \tag{8}$$

$$sensibilidad = \frac{TP}{TP + FN} \tag{9}$$

$$exactitud = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

$$\text{puntaje F1} = 2 \cdot \frac{\text{precision} \cdot \text{sensibilidad}}{\text{precision} + \text{sensibilidad}} \tag{11}$$

IV.2. Línea base de comparación

Para llevar a cabo la evaluación experimental, se utiliza una línea base de comparación en los experimentos, considerando un modelo que siempre arroja como resultado a la clase mayoritaria. Este valor se calcula utilizando el $rating$ con mayor frecuencia en el corpus, que en este caso es de 5 (ver sección III.1). En la tabla II de la sección III.1, se observa que aproximadamente el 30 % de las instancias de corpus tienen una calificación igual a 4, mientras que la calificación restante (de 1 a 3) representa casi el 20 %. Considerando un modelo que siempre otorga soluciones de la clase mayoritaria, se obtiene una exactitud del 54.1854 % (ver tabla VIII).

Tabla VIII. Resultados del modelo de clase mayoritaria

MAE	**MSE**	**RMSE**	**exactitud**	**puntaje F1**
0.72246	1.49779	1.22384	0.54185	0.14057

IV.3. Resultados del filtrado colaborativo

La tabla IX muestra el desempeño del modelo del filtrado colaborativo bajo el enfoque basado en usuarios. Cada valor de k representa la cantidad de vecinos que se toma en cuenta para tomar una decisión. Se observa que el mejor resultado se obtiene cuando k = 10, con los menores valores de error y con la mayor exactitud. Sin embargo, esto no fue el caso del puntaje F1, donde el valor más alto resulta cuando k = 25. Se tiene en cuenta que no hay una diferencia significativa con respecto a los resultados restantes. Con respecto a la línea de base de la clase mayoritaria, este enfoque obtiene

Tabla IX. Desempeño del modelo colaborativo bajo el enfoque basado en usuarios.

Valor de k	**MAE**	**MSE**	**RMSE**	**exactitud**	**puntaje F1**
10	**0.79083**	**1.07988**	**1.03917**	**0.32599**	0.13035
20	0.79374	1.0937	1.0458	0.32452	0.12993
25	0.79348	1.09309	1.04551	0.32745	**0.13188**
30	0.79354	1.09302	1.04548	0.32599	0.13091
35	0.79359	1.09314	1.04553	0.32599	0.13091

valores más bajos de MSE y RMSE; sin embargo, no supera al resto de las métricas. Así, aunque el modelo de recomendación colaborativo basado en usuarios es aceptable al no adquirir errores promedio, tiene mayor sesgo de error MAE que la línea base de comparación.

La tabla X muestra el desempeño del modelo del filtrado colaborativo bajo el enfoque basado en ítems. A diferencia del enfoque basado en usuarios, el vecindario en k representa la cantidad de lugares. Dado que solamente existen 18 ítems (ver tabla Xde los lugares turísticos), entonces se utiliza un valor de k menor a ese. El resultado con las medidas de error más bajas y la mayor exactitud es cuando k = 3. El mejor puntaje F1 se obtiene cuando k = 1. De manera similar al enfoque basado en usuarios, no hay diferencias significativas con respecto a los resultados restantes. Además, este enfoque también obtiene valores de MSE y RMSE más bajos que los de la línea base de comparación, pero no supera al resto de las métricas.

Tabla X. Desempeño del modelo colaborativo bajo el enfoque basado en ítems.

Valor de k	**MAE**	**MSE**	**RMSE**	**exactitud**	**puntaje F1**
1	0.79588	1.11894	1.0578	0.32599	**0.13167**
3	**0.78267**	**1.07929**	**1.03889**	**0.32892**	0.12528
5	0.79882	1.11894	1.0578	0.32158	0.12819
7	0.79735	1.11747	1.05711	0.32305	0.12969
9	0.78120	1.07489	1.036768	0.328928	0.12517

Tabla XI. Desempeño de los modelos basados en el filtrado demográfico.

Modelo	**MAE**	**MSE**	**RMSE**	**exactitud**	**puntaje F1**
KNN 10	0.69456	1.33186	1.15406	0.51982	0.19155
KNN 20	0.70778	1.39207	1.17986	0.52569	0.18355
KNN 25	0.70044	1.37885	1.17424	0.53010	0.18286
KNN 30	0.69603	1.36857	1.16986	0.53303	0.18070
KNN 35	0.68428	1.3392	1.15724	0.53597	0.18438
RF	**0.66666**	**1.29809**	**1.13933**	**0.54478**	**0.20378**
NN	0.68428	1.3392	1.15724	0.54185	0.17582

IV.4. Resultados del filtrado demográfico

La tabla XI muestra el rendimiento de los modelos que utilizan el filtrado demográfico, entrenados por los siguientes algoritmos de aprendizaje: árboles aleatorios (o RF, $random\ forest$), redes neuronales (o NN, $neural\ networks$) y KNN con k = $10, 20, 25, 30, 35$ (ver sección III.2.4). De la tabla se observa que el mejor resultado se obtiene del modelo entrenado por RF, arrojando los valores de error (MAE, MSE y RMSE) más pequeños y el mejor desempeño tanto en exactitud como en el puntaje F1. A diferencia de los modelos de recomendación colaborativos, todos los modelos generados a través del filtrado demográfico superan la línea base de comparación en todas las métricas de evaluación. Esto significa que esta técnica de filtrado, independientemente del algoritmo de aprendizaje utilizado, fue más eficiente.

IV.5. Resultados generales

Los resultados experimentales muestran que los modelos basados en el filtrado demográfico superan la línea base de comparación. Sin embargo, se obtienen mejores valores de error MSE y RMSE en los modelos de filtrado colaborativo.

Tanto MAE como RMSE expresan el error promedio de predicción del modelo en unidades de la variable de interés. Ambas métricas pueden oscilar entre cero e infinito y son indiferentes a la dirección de los errores. Son puntuaciones de orientación negativa, lo que significa que los valores más bajos son mejores. RMSE tiene la ventaja de penalizar más los errores grandes, por lo que puede ser más apropiado en los casos donde el filtrado demográfico tiene menos sesgo promedio del modelo, y los filtrados colaborativos por usuarios y por ítems tiene menos errores grandes.

Capítulo V

Conclusiones y trabajo futuro

En este trabajo de tesis se proponen tres sistemas de recomendación turística: dos basados en el filtrado colaborativo y uno en el filtrado demográfico. Para la generación de estos sistemas de recomendación, se construyó un corpus donde se recolectaron un total de 2,263 opiniones y $ratings$ del sitio TripAdvisor.com. Estas opiniones y $ratings$ fueron vertidas por 2,033 turistas acerca de 18 sitios turísticos Nayarit, México. Hasta donde el autor de la presente tesis conoce, éste sería el primer corpus con reseñas de turismo en español diseñado para el entrenamiento de sistemas de recomendación turística.

Los resultados experimentales muestran que el enfoque de filtrado basado en datos demográficos supera a la clase mayoritaria (utilizada como referencia) en todas las métricas de evaluación. Este no es el caso de los enfoques de filtrado basados en la colaboración (usuarios e ítems), aunque obtuvieron los valores generales más bajos de MSE y RMSE. Cabe enfatizar que la técnica de filtrado basado en datos demográficos obtuvo una mayor eficiencia independientemente del algoritmo de aprendizaje utilizado.

V.1. Trabajo futuro

Como trabajo futuro, sería interesante explorar otros enfoques de recomendación, como los modelos basados en el contexto u otros que utilizan el aprendizaje profundo y el procesamiento del lenguaje natural. También resulta interesante escalar el corpus para abarcar otros lugares turísticos en México y con una mayor diversidad de tipología turística.

Referencias bibliográficas

Adomavicius, G. y Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge & Data Engineering*, (6): 734–749.

Adomavicius, G. y Tuzhilin, A. (2011). Context-aware recommender systems. En *Recommender systems handbook*, pp. 217–253. Springer.

Aguirre Quezada, J. P. (2020). Caída del turismo por la covid-19, desafío para México y experiencias internacionales. `http://bibliodigitalibd.senado.gob.mx/handle/123456789/4882`. Último acceso: 17 Nov de 2021.

Al-Ghobari, M., Muneer, A., y Fati, S. M. (2021). Location-aware personalized traveler recommender system (lapta) using collaborative filtering knn. *Computers Material & Continua*, **68**.

Amatriain, X. (2014). Recommender systems. `https://cirocavani.wordpress.com/2014/08/06/recommender-systems-machine-learning-summer-school-2014-cmu/`. Machine Learning Summer School. Último acceso: 17 Nov de 2021.

Arce Cardenas, S., Fajardo Delgado, D., y Carmona, M. Á. Á. (2020). Tecnm at mex-a3t 2020: Fake news and aggressiveness analysis in mexican spanish. En *IberLEF@ SEPLN*, pp. 265–272.

Argamon, S., Koppel, M., Pennebaker, J. W., y Schler, J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM*, **52**(2): 119–123.

Ashraf, S., Iqbal, H. R., y Nawab, R. M. A. (2016). Cross-genre author profile prediction using stylometry-based approach. En *CLEF (Working Notes)*, pp. 992–999.

Bilal, M., Gani, A., Lali, M. I. U., Marjani, M., y Malik, N. (2019). Social profiling: A review, taxonomy, and challenges. *Cyberpsychology, Behavior, and Social Networking*, **22**(7): 433–450.

Billsus, D. y Pazzani, M. J. (2000). User modeling for adaptive news access. *User modeling and user-adapted interaction*, **10**(2-3): 147–180.

Binucci, C., De Luca, F., Di Giacomo, E., Liotta, G., y Montecchiani, F. (2017). Designing the content analyzer of a travel recommender system. *Expert Systems with Applications*, **87**: 199–208.

Blei, D. M., Ng, A. Y., y Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of Machine Learning Research*, **3**: 993–1022.

Breese, J. S., Heckerman, D., y Kadie, C. (2013). Empirical analysis of predictive algorithms for collaborative filtering. *arXiv preprint arXiv:1301.7363*.

Burke, R. (2002). Hybrid recommender systems: Survey and experiments: In journal user modeling and user-adapted interaction, vol 12, issue 4.

EFE, A. (2020). Estiman caída del 10% en el pib turístico de méxico. https://www.efe.com/efe/america/mexico/estiman-caida-del-10-en-el-pib-tu-ristico-de-mexico/50000545-4233506. Último acceso: 17 Nov de 2021.

El economista (2019). Sector de viajes y turismo creció más que el pib. https://www.eleconomista.com.mx/empresas/Sector-de-viajes-y-turismo-crecio-mas-que-el-PIB--20190301-0003.html. Último acceso: 17 Nov de 2021.

Fararni, K. A., Nafis, F., Aghoutane, B., Yahyaouy, A., Ri□, J., y Sabri, A. (2021). Hybrid recommender system for tourism based on big data and ai: A conceptual framework. *Big Data Mining and Analytics*, **4**(1): 47–55.

Gelbukh, A. y Sidorov, G. (2006). *Procesamiento automático del español con enfoque en recursos léxicos grandes*. ISBN 970-36-0264-9.

Ghazanfar, M. A. y Prugel-Bennett, A. (2010). A scalable, accurate hybrid recommender system. En *2010 Third International Conference on Knowledge Discovery and Data Mining*, pp. 94–98. IEEE.

González, O. E. y Jacques, S. M. (2017). Estado del arte en los sistemas de recomendación. *Research in Computing Science*, **135**: 25–40.

Grčar, M., Fortuna, B., Mladenič, D., y Grobelnik, M. (2006). Knn versus svm in the collaborative filtering framework. En *Data Science and Classification*, pp. 251–260. Springer.

Hamid, R. A., Albahri, A., Alwan, J. K., Al-qaysi, Z., Albahri, O., Zaidan, A., Alnoor, A., Alamoodi, A., y Zaidan, B. (2021). How smart is e-tourism? a systematic review of smart tourism recommendation system applying data management. *Computer Science Review*, **39**: 100337.

Han, J. y Lee, H. (2015). Adaptive landmark recommendations for travel planning: Personalizing and clustering landmarks using geo-tagged social media. *Pervasive and Mobile Computing*, **18**: 4 – 17.

Hedlund, J. y Nilsson Tengstrand, E. (2020). A comparison between di□erent recommender system approaches for a book and an author recommender system. http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-166378. Tesis de maestría. Linköping University, Department of Computer and Information Science.

Hofmann, T. (2003). Collaborative filtering via gaussian probabilistic latent semantic analysis. En *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 259–266.

INEGI (2019). Estadísticas a propósito del día mundial del turismo. `https://www.inegi.org.mx/contenidos/saladeprensa/aproposito/2019/turismo2019_Nal.pdf`. Último acceso: 17 Nov de 2021.

Kandias, M., Mitrou, L., Stavrou, V., y Gritzalis, D. (2017). Profiling online social networks users: an omniopticon tool. *International Journal of Social Network Mining*, **2**(4): 293–313.

Kuanr, M. y Mohanty, S. N. (2020). Location-based personalised recommendation systems for the tourists in india. *International Journal of Business Intelligence and Data Mining*, **17**(3): 377–392.

Kzaz, L., Dakhchoune, D., y Dahab, D. (2018). Tourism recommender systems: an overview of recommendation approaches. *International Journal of Computer Applications*, **975**: 8887.

Li, Q., Li, S., Zhang, S., Hu, J., y Hu, J. (2019). A review of text corpus-based tourism big data mining. *Applied Sciences*, **9**(16).

Logesh, R., Subramaniyaswamy, V., Vijayakumar, V., y Li, X. (2019). E□cient user profiling based intelligent travel recommender system for individual and group of users. *Mobile Networks and Applications*, **24**(3): 1018–1033.

Menk, A., Sebastia, L., y Ferreira, R. (2017). Curumim: A serendipitous recommender system for tourism based on human curiosity. En *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 788–795, Nov.

Menk, A., Sebastia, L., y Ferreira, R. (2019). Recommendation systems for tourism based on social networks: A survey.

Rangel, F., Rosso, P., Montes-y Gómez, M., Potthast, M., y Stein, B. (2018). Overview of the 6th author profiling task at pan 2018: multimodal gender identification in twitter. *Working Notes Papers of the CLEF*.

Ricci, F., Rokach, L., y Shapira, B. (2011). Introduction to recommender systems handbook. En *Recommender systems handbook*, pp. 1–35. Springer.

Salakhutdinov, R., Mnih, A., y Hinton, G. (2007). Restricted boltzmann machines for collaborative filtering. En *Proceedings of the 24th international conference on Machine learning*, pp. 791–798.

Salunke, S. S. (2014). *Selenium webdriver in Python: Learn with examples*, Vol. 70. CreateSpace Independent Publishing Platform, USA,.

Schafer, J. B., Frankowski, D., Herlocker, J., y Sen, S. (2007). Collaborative filtering recommender systems. En *The adaptive web*, pp. 291–324. Springer.

SECTUR (2018). Ranking mundial de turismo internacional. `https://www.datatur.sectur.gob.mx/SitePages/RankingOMT.aspx`. Último acceso: 17 Nov de 2021.

Shen, J., Deng, C., y Gao, X. (2016). Attraction recommendation: Towards personalized tourism via collective intelligence. *Neurocomputing*, **173**: 789–798.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, **60**(3): 538–556.

Vu, H. Q., Li, G., Law, R., y Zhang, Y. (2019). Exploring tourist dining preferences based on restaurant reviews. *Journal of Travel Research*, **58**(1): 149–167.

Welle, D. (2020). El impacto al turismo arrastrará a la economía mexicana. `https://www.dw.com/es/el-impacto-al-turismo-arrastra-a-la-econom%C3%ADa-mexicana/a-53137428`. Último acceso: 17 Nov de 2021.

Xiong, H., Zhou, Y., Hu, C., Wei, X., y Li, L. (2017). A novel recommendation algorithm frame for tourist spots based on multi-clustering bipartite graphs. En *2017 IEEE 2nd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, pp. 276–282. IEEE.

Yochum, P., Chang, L., Gu, T., y Zhu, M. (2020). Linked open data in location-based recommendation system on tourism domain: A survey. *IEEE Access*, **8**: 16409–16439.

Young, T., Hazarika, D., Poria, S., y Cambria, E. (2018). Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine*, **13**(3): 55–75.

Zafarani, R., Abbasi, M. A., y Liu, H. (2014). *Social media mining: an introduction*. Cambridge University Press. ISBN 978-1-107-01885-3.

Zitnick, C. y Kanade, T. (2004). Maximum entropy for collaborative filtering. 20th conf. on uncertainty in artificial intelligence (pp. 636-643). arlington, virginia.

Apéndices

Apéndice A.

Trabajo alterno MEX-A3T

Derivado de los objetivos de la presente tesis, se participó en en el trabajo del MEX-A3T 2020: *Fake news and offensive language detection in Mexican Spanish* como parte del *Iberian Languages Evaluation Forum* (IberLEF 2020). Este trabajo consistió en dos tareas principales: la identificación de noticias falsas (*FakeNews*) y un análisis de agresividad (*Aggressiveness*) para el perfilado de autores. Para abordar estas dos tareas, en el evento se planteó una competencia donde participaron diversas universidades y centros de investigación científica en México. En el presente trabajo, se propuso una solución para las dos tareas que utiliza técnicas de procesamiento de lenguaje natural (PLN) y del perfilado de autores. En este apéndice se describe, además de los conceptos del PLN y del perfilado de autores, la metodología que se utilizó para resolver las dos tareas y los resultados obtenidos.

A.1. Procesamiento del Lenguaje Natural

El *procesamiento del lenguaje natural* (PLN) es una gama de técnicas computacionales para el análisis automático y la representación del lenguaje humano (Young *et al.*, 2018). El PLN abarca diversas tareas como desambiguación de palabras, identificación de colocaciones, etiquetado gramatical, traducción automática, identificación de enti-

dades nombradas, agrupamiento y clasificación de textos (Gelbukh y Sidorov, 2006).

Las tareas de clasificación y el agrupamiento de textos son comúnmente abordadas por medio de aprendizaje máquina supervisado, mientras que el agrupamiento generalmente se trata mediante aprendizaje no supervisado. Algunas aplicaciones de la clasificación y el agrupamiento automático de textos son las siguientes: análisis de sentimientos, identificación de idioma, identificación de género literario, agrupamiento por temática, entre otros.

Por otro lado, para la clasificación automática de textos, es necesario cuantificar y pasar a un espacio vectorial las características de los objetos de estudio. En el caso de la clasificación automática de textos las características son el texto en sí, siendo posible separarlas en dos grupos: características de contenido y características estilísticas (Argamon *et al.*, 2009).

A.2. Perfilado de autores

El *perfilado de autores* es el análisis de texto para predecir diferentes atributos de los autores del contenido, tales como género, edad, personalidad, idioma nativo u orientación política, entre otros (Rangel *et al.*, 2018). El perfilado de autores considera un conjunto de características estilizadas para identificar el estilo de escritura de los autores. Normalmente, el estilo de un autor puede clasificarse en tres tipos de características: léxicas, sintácticas y basadas en caracteres. Las *características léxicas* representan al texto como una secuencia de fichas o tokens que forman oraciones,

párrafos y documentos. Un $token$ puede ser un número, una letra o una palabra alfabética de un signo de puntuación. Estos tokens se utilizan para obtener estadísticas como la longitud promedio de las oraciones y la longitud promedio de las palabras. Estas características tienen la capacidad de obtener información de un texto en cualquier idioma sin requisitos especiales (Stamatatos, 2009). Las $características\ sintácticas$ consisten en palabras de función y etiquetas de partes del discurso, donde el patrón sintáctico puede varíar significativamente de un autor a otro (Ashraf $et\ al.$, 2016). Las $características\ basadas\ en\ caracteres$ consideran el texto como una secuencia de caracteres que incluyen en su medición el recuento de caracteres, signos de puntuación y dígitos (Stamatatos, 2009).

A.3. Metodología propuesta

La metodología propuesta para perfilado de autores en las tareas de identificación de $FakeNews$ y de $Aggressiveness$ consta de los siguientes tres pasos: preprocesamiento del texto, representación del texto y construcción de los modelos de clasificación.

El preprocesamiento de texto es comúnmente el primer paso de un sistema de procesamiento de lenguaje natural e incluye un conjunto de técnicas diseñadas para transformar documentos de texto en una forma de representación adecuada para el procesamiento automático. Las técnicas de preprocesamiento que se emplearon en este trabajo incluyeron: el uso de expresiones regulares, la $tokenización$, la eliminación de puntuación, símbolos, palabras vacías y $stemming$. Las expresiones regulares permiten

identificar algunas palabras incorrectas para el español mexicano, principalmente aquellas en las que la misma vocal aparece posteriormente tres veces o más. La mejor manera de hacer esto fue empleando la biblioteca “re” en Python. También se utilizó el kit de herramientas de lenguaje natural (NLTK) para realizar la $tokenización$, dividiendo los textos en palabras como elementos esenciales. Durante este proceso, también se eliminaron los signos de puntuación, los caracteres especiales o símbolos, así como las palabras vacías innecesarias como “el”, “la”, “los”. Posteriormente, se utilizó la biblioteca de raíces de $Snowball$ para reducir las palabras derivadas a su forma o raíz original realizando el truncamiento de sufijos. Finalmente, para reducir aún más el número de palabras sin significado, se ignoraron aquellas que aparecen menos de 20 o 40 veces.

Después del preprocesamiento del texto, se buscó identificar el conjunto de palabras que mejor describen el contexto textual. Extraer estas palabras, también llamadas términos o palabras clave, es el proceso para asignar un valor numérico que representa la relevancia de cada palabra con respecto a las demás dentro del corpus. En particular, se utilizaron dos métodos basados en un enfoque estadístico simple: el término-frecuencia (TF) y el término-frecuencia inversa de documentos (TF-IDF). TF define la importancia local que tiene cada término en un documento en función de su frecuencia; es decir, si una palabra w aparece con frecuencia en un documento, entonces lo más importante es w. IDF captura cuántos documentos aparece una palabra con respecto al número total de palabras en el corpus, es decir, resalta la rareza de la palabra. Se usaron las implementaciones de TF y TF-IDF incluidas en la biblioteca scikit-learn.

Finalmente, para construir los modelos de clasificación, se usaron los siguientes algoritmos de aprendizaje automático implementados en scikit-learn: los k-vecinos más cercanos (KNN) por $k = 3, 7, 11$, la máquina de vectores de soporte (SVM) con núcleos de función de base lineal y radial (RBF), árboles de decisión (DT), red neuronal (NN) e ingenuo Bayes (NB). Se generaron estos modelos utilizando el conjunto de entrenamiento utilizando una validación cruzada de 10 veces.

A.4. Resultados

Para la tarea de perfilado de autores, se dividió el conjunto de datos en 10 tomando el primer subconjunto como validación y los otros como entrenamiento. Posteriormente, se toma el segundo subconjunto como validación y el resto para entrenamiento, y así consecutivamente. Se repitió este proceso hasta que cada subconjunto se haya utilizado para la validación. Finalmente, se agregaron las matrices de confusión y se analizaron los resultados.

Las tablas XII y XIII muestran el rendimiento de los modelos de clasificación propuestos aplicados al conjunto de datos de $fake\ news$ utilizando los métodos TF y TF-IDF, respectivamente. El mejor resultado para este conjunto de datos es la combinación de NN sin utilizar las técnicas de palabras vacías y derivación, e independientemente del uso de TF y TF-IDF. Cabe resaltar que, a excepción de la SVM con RBF, existe una diferencia notable entre los resultados de NN con respecto al resto. También se observa que, en general, los resultados son ligeramente mejores cuando se usa TF-IDF

que con TF.

Por otro lado, las tablas XIV y XV muestran el rendimiento de los modelos de clasificación propuestos aplicados al conjunto de datos de $Aggressiveness$ utilizando los métodos TF y TF-IDF, respectivamente. El mejor resultado para este conjunto de datos es mediante la combinación de NN con el método TF-IDF y sin utilizar las técnicas de palabras vacías y derivación. Al igual que el conjunto de datos de noticias falsas, los resultados del conjunto de datos de $agresividad$ son ligeramente mejores cuando se usa TF-IDF que con TF. Por otro lado, y a diferencia de los resultados de la clasificación de noticias falsas, el mejor modelo que utiliza el método TF es el SVM con RBF.

Todos estos resultados se obtuvieron ignorando las palabras que se repiten menos de 20 veces para ambos conjuntos de datos (tablas XII-XV). Por otro lado, el nuevo conjunto de datos incluye, además del texto completo de la noticia, un encabezado que describe el título de la noticia. Se realizaron experimentos considerando el encabezado y no considerándolo. Las tablas XII y XIII muestran sólo los resultados cuando no se considera el encabezado, ya que estos presentan mejores resultados.

Finalmente, para ambos conjuntos de datos, los mejores resultados se obtuvieron conservando las palabras vacías y omitiendo el proceso de vaporización. Se conjetura que considerar tales palabras para estos casos particulares puede distinguir las clases ($Aggressiveness$ / $FakeNews$) en los textos.

Tabla XII. Resultados de rendimiento de los modelos propuestos para el conjunto de datos de *FakeNews* utilizando TF en la etapa de validación.

clasificador	exactitud	precisión	sensibilidad	F1	*Stopwords*	*Stemming*
KNN_3	63.265	0.664±0.069	0.631±0.223	0.613±0.088	Si	Si
KNN_7	62.48	0.668±0.083	0.623±0.259	0.597±0.106	Si	Si
KNN_11	61.538	0.660±0.082	0.613±0.268	0.584±0.114	Si	Si
L_SVM	50.392	0.252±0.252	0.500±0.500	0.335±0.335	Si	Si
RBF SVM	74.411	0.744±0.004	0.744±0.003	0.744±0.001	Si	Si
DT	65.62	0.657±0.012	0.656±0.027	0.656±0.008	Si	Si
NN	76.766	0.768±0.000	0.768±0.005	0.768±0.003	Si	Si
NB	65.777	0.658±0.005	0.658±0.004	0.658±0.001	Si	Si
KNN_3	62.48	0.670±0.085	0.623±0.262	0.596±0.108	No	Si
KNN_7	63.108	0.699±0.115	0.629±0.296	0.594±0.122	No	Si
KNN_11	60.911	0.676±0.106	0.607±0.312	0.565±0.138	No	Si
L_SVM	50.392	0.252±0.252	0.500±0.500	0.335±0.335	No	Si
RBF SVM	75.039	0.750±0.006	0.750±0.006	0.750±0.000	No	Si
DT	69.231	0.693±0.010	0.692±0.016	0.692±0.003	No	Si
NN	75.51	0.755±0.002	0.755±0.002	0.755±0.002	No	Si
NB	66.091	0.661±0.001	0.661±0.009	0.661±0.005	No	Si
KNN_3	59.812	0.604±0.022	0.597±0.126	0.591±0.053	Si	No
KNN_7	61.381	0.625±0.036	0.613±0.157	0.603±0.064	Si	No
KNN_11	62.794	0.649±0.054	0.626±0.193	0.613±0.077	Si	No
L_SVM	50.392	0.252±0.252	0.500±0.500	0.335±0.335	Si	No
RBF SVM	74.568	0.746±0.016	0.746±0.026	0.746±0.005	Si	No
DT	60.283	0.627±0.058	0.604±0.212	0.585±0.086	Si	No
NN	76.138	0.762±0.010	0.761±0.014	0.761±0.002	Si	No
NB	72.841	0.729±0.010	0.728±0.029	0.728±0.009	Si	No
KNN_3	61.695	0.700±0.127	0.614±0.326	0.570±0.143	No	No
KNN_7	58.556	0.667±0.115	0.583±0.355	0.524±0.171	No	No
KNN_11	59.969	0.709±0.150	0.597±0.366	0.536±0.172	No	No
L_SVM	50.392	0.252±0.252	0.500±0.500	0.335±0.335	No	No
RBF SVM	78.493	0.788±0.026	0.785±0.050	0.784±0.012	No	No
DT	66.876	0.669±0.002	0.669±0.004	0.669±0.003	No	No
NN	**79.121**	**0.792**±0.012	**0.791**±0.025	**0.791**±0.007	No	No
NB	75.981	0.760±0.004	0.760±0.013	0.760±0.005	No	No

Tabla XIII. Resultados de rendimiento de los modelos propuestos para el conjunto de datos de *FakeNews* utilizando TF-IDF en la etapa de validación.

clasificador	exactitud	precisión	sensibilidad	F1	*Stopwords*	*Stemming*
KNN_3	57.614	0.585±0.025	0.575±0.170	0.563±0.076	Si	Si
KNN_7	60.597	0.631±0.055	0.604±0.224	0.584±0.095	Si	Si
KNN_11	62.009	0.651±0.067	0.618±0.232	0.597±0.095	Si	Si
L_SVM	50.392	0.252±0.252	0.500±0.500	0.335±0.335	Si	Si
RBF SVM	76.138	0.762±0.013	0.762±0.020	0.761±0.003	Si	Si
DT	65.62	0.658±0.014	0.656±0.055	0.655±0.021	Si	Si
NN	77.237	0.772±0.004	0.772±0.003	0.772±0.001	Si	Si
NB	65.777	0.658±0.002	0.658±0.006	0.658±0.004	Si	Si
KNN_3	60.597	0.626±0.048	0.604±0.206	0.588±0.087	No	Si
KNN_7	60.597	0.639±0.066	0.604±0.250	0.579±0.107	No	Si
KNN_11	62.951	0.672±0.084	0.628±0.254	0.603±0.103	No	Si
L_SVM	50.392	0.252±0.252	0.500±0.500	0.335±0.335	No	Si
RBF SVM	76.138	0.762±0.010	0.761±0.014	0.761±0.002	No	Si
DT	63.265	0.633±0.001	0.633±0.019	0.632±0.009	No	Si
NN	78.022	0.780±0.003	0.780±0.001	0.780±0.001	No	Si
NB	66.719	0.667±0.001	0.667±0.009	0.667±0.005	No	Si
KNN_3	62.951	0.633±0.021	0.629±0.091	0.626±0.036	Si	No
KNN_7	63.108	0.635±0.020	0.630±0.089	0.628±0.035	Si	No
KNN_11	63.579	0.645±0.036	0.635±0.135	0.629±0.052	Si	No
L_SVM	50.392	0.252±0.252	0.500±0.500	0.335±0.335	Si	No
RBF SVM	74.882	0.751±0.024	0.749±0.042	0.748±0.009	Si	No
DT	63.736	0.663±0.067	0.639±0.193	0.624±0.071	Si	No
NN	76.609	0.766±0.006	0.766±0.006	0.766±0.000	Si	No
NB	73.155	0.732±0.007	0.731±0.023	0.731±0.008	Si	No
KNN_3	58.713	0.663±0.110	0.584±0.347	0.529±0.166	No	No
KNN_7	55.102	0.640±0.110	0.548±0.405	0.460±0.221	No	No
KNN_11	54.474	0.668±0.142	0.541±0.434	0.437±0.247	No	No
L_SVM	50.392	0.252±0.252	0.500±0.500	0.335±0.335	No	No
RBF SVM	78.65	0.787±0.011	0.787±0.014	0.786±0.002	No	No
DT	67.19	0.672±0.006	0.672±0.008	0.672±0.001	No	No
NN	**81.476**	**0.815**±0.008	**0.815**±0.017	**0.815**±0.004	No	No
NB	74.568	0.746±0.005	0.746±0.004	0.746±0.000	No	No

Tabla XIV. Resultados de rendimiento de los modelos propuestos para el conjunto de datos de *Aggressiveness* mediante el uso de TF en la etapa de validación.

Clasificador	exactitud	Precisión	sencibilidad	F1	*Stopwords*	*Stemming*
KNN_3	75.762	0.713±0.067	0.641±0.278	0.654±0.190	Si	Si
KNN_7	75.99	0.738±0.028	0.621±0.331	0.631±0.218	Si	Si
KNN_11	75.686	0.741±0.020	0.611±0.348	0.617±0.232	Si	Si
L_SVM	72.031	0.806±0.088	0.519±0.479	0.456±0.380	Si	Si
RBF SVM	81.344	0.811±0.004	0.711±0.243	0.736±0.143	Si	Si
DT	78.735	0.768±0.029	0.676±0.264	0.696±0.167	Si	Si
NN	81.435	0.805±0.015	0.718±0.228	0.742±0.137	Si	Si
NB	60.413	0.641±0.232	0.667±0.149	0.597±0.053	Si	Si
KNN_3	75.171	0.704±0.069	0.629±0.292	0.640±0.200	No	Si
KNN_7	75.034	0.718±0.042	0.607±0.341	0.613±0.231	No	Si
KNN_11	75.049	0.727±0.030	0.601±0.355	0.604±0.241	No	Si
L_SVM	71.53	0.809±0.095	0.510±0.490	0.436±0.397	No	Si
RBF SVM	81.556	0.817±0.002	0.713±0.245	0.738±0.143	No	Si
DT	78.553	0.767±0.028	0.672±0.270	0.691±0.170	No	Si
NN	81.283	0.808±0.008	0.712±0.240	0.736±0.142	No	Si
NB	62.096	0.648±0.228	0.677±0.134	0.612±0.058	No	Si
KNN_3	77.825	0.735±0.075	0.691±0.208	0.705±0.147	Si	No
KNN_7	77.886	0.744±0.055	0.676±0.244	0.693±0.162	Si	No
KNN_11	78.083	0.761±0.028	0.663±0.279	0.682±0.178	Si	No
L_SVM	73.199	0.807±0.080	0.541±0.455	0.499±0.342	Si	No
RBF SVM	81.116	0.796±0.025	0.718±0.222	0.740±0.136	Si	No
DT	77.158	0.775±0.005	0.631±0.335	0.643±0.215	Si	No
NN	80.965	0.797±0.019	0.712±0.232	0.735±0.141	Si	No
NB	63.325	0.650±0.221	0.681±0.113	0.622±0.065	Si	No
KNN_3	74.867	0.692±0.091	0.644±0.250	0.655±0.179	No	No
KNN_7	76.399	0.727±0.054	0.645±0.283	0.659±0.189	No	No
KNN_11	76.824	0.741±0.037	0.643±0.297	0.658±0.194	No	No
L_SVM	71.045	0.730±0.020	0.501±0.499	0.417±0.414	No	No
RBF SVM	**81.753**	**0.817**±0.001	**0.717**±0.239	**0.742**±0.139	No	No
DT	77.977	0.785±0.007	0.645±0.319	0.662±0.200	No	No
NN	81.435	0.809±0.008	0.715±0.236	0.739±0.140	No	No
NB	64.22	0.656±0.220	0.689±0.112	0.631±0.066	No	No

Tabla XV. Resultados de rendimiento de los modelos propuestos para el conjunto de datos de *Aggressiveness* mediante el uso de TF-IDF en la etapa de validación.

Clasificador	exactitud	Precisión	sensibilidad	F1	*Stopwords*	*Stemming*
KNN_3	74.063	0.692±0.065	0.598±0.339	0.602±0.235	Si	Si
KNN_7	74.215	0.725±0.021	0.579±0.388	0.571±0.271	Si	Si
KNN_11	73.639	0.728±0.010	0.563±0.413	0.544±0.297	Si	Si
L_SVM	71.728	0.838±0.123	0.513±0.487	0.442±0.392	Si	Si
RBF SVM	80.98	0.813±0.004	0.701±0.258	0.726±0.151	Si	Si
DT	78.538	0.765±0.030	0.673±0.267	0.692±0.169	Si	Si
NN	81.283	0.806±0.011	0.714±0.236	0.737±0.141	Si	Si
NB	60.458	0.639±0.231	0.665±0.144	0.597±0.055	Si	Si
KNN_3	73.881	0.701±0.047	0.582±0.372	0.578±0.260	No	Si
KNN_7	73.487	0.736±0.001	0.557±0.424	0.532±0.308	No	Si
KNN_11	73.047	0.726±0.005	0.548±0.434	0.518±0.321	No	Si
L_SVM	71.455	0.826±0.113	0.508±0.492	0.432±0.400	No	Si
RBF SVM	81.42	0.821±0.010	0.707±0.256	0.732±0.148	No	Si
DT	78.492	0.766±0.028	0.671±0.270	0.690±0.171	No	Si
NN	81.541	0.806±0.015	0.720±0.227	0.743±0.136	No	Si
NB	62.051	0.650±0.229	0.679±0.139	0.612±0.057	No	Si
KNN_3	76.596	0.719±0.078	0.669±0.232	0.683±0.163	Si	No
KNN_7	77.233	0.745±0.039	0.652±0.285	0.669±0.185	Si	No
KNN_11	77.582	0.776±0.000	0.641±0.322	0.655±0.204	Si	No
L_SVM	73.093	0.823±0.097	0.538±0.459	0.493±0.348	Si	No
RBF SVM	81.132	0.798±0.021	0.716±0.227	0.738±0.138	Si	No
DT	76.885	0.772±0.004	0.626±0.341	0.636±0.219	Si	No
NN	81.04	0.797±0.021	0.715±0.228	0.737±0.139	Si	No
NB	63.598	0.648±0.218	0.679±0.103	0.623±0.069	Si	No
KNN_3	76.96	0.741±0.041	0.648±0.288	0.664±0.189	No	No
KNN_7	78.508	0.775±0.014	0.664±0.288	0.683±0.180	No	No
KNN_11	79.478	0.797±0.003	0.675±0.286	0.696±0.173	No	No
L_SVM	71.318	0.833±0.121	0.505±0.494	0.427±0.405	No	No
RBF SVM	82.345	0.827±0.006	0.725±0.235	0.751±0.134	No	No
DT	78.159	0.790±0.011	0.647±0.319	0.664±0.199	No	No
NN	**82.345**	**0.820**±0.005	**0.730**±0.223	**0.754**±0.130	No	No
NB	65.418	0.659±0.214	0.693±0.094	0.640±0.071	No	No

Apéndice B

Productos académicos

B.1. Producto derivado de la tesis

Se aceptó el trabajo titulado "*A Tourist Recommendation Systems: A Study Case in Mexico*" para presentarse en el *20th Mexican International Conference on Artificial Intelligence* (MICAI 2020), el cual se llevará a cabo del 25 al 30 de Octubre del presente año. Los resultados se publicarán en el *Springer Lectures Notes in Artificial Intelligence*, indizado en el *Web of Science*, Scopus y EI.

A Tourist Recommendation System: A Study Case in Mexico

Samuel Arce-Cardenas[1][0000−0002−2547−0047],
Daniel Fajardo-Delgado[1][0000−0001−8215−5927],
Miguel Á. Álvarez-Carmona[2](✉)[0000−0003−4421−5575], and
Juan Pablo Ramírez-Silva[3][0000−0001−6391−9143]

[1] Tecnológico Nacional de México, Av. Tecnológico 100, Guzman City, 49100, Mexico
{m19291003, daniel.fd}cdguzman.tecnm.mx
[2] CICESE-UT3, Andador 10 Ciudad del Conocimiento, Tepic, 63173, Mexico
malvarez@cicese.edu.mx
[3] Universidad Autónoma de Nayarit, Ciudad de la Cultura, Tepic, 63155, Mexico
pablor@uan.edu.mx

Abstract. The present work deals with implementing tourist recommendation systems designed to predict the user preferences about a place or tourist activity in Mexico. Three recommendation systems have been proposed: two based on collaborative filtering (user and items) and the other based on demographic issues. To this aim, a corpus has been built by collecting 2,263 ratings from TripAdvisor.com about eighteen tourist places in Mexico. Experimental results show that the demographic-based recommendation system outperforms those based on collaborative filtering, obtaining a mean absolute error of 0.67 and a mean square error of 1.2980. These results also show significant improvement over a majority class baseline based on a sizeable unbalanced corpus.

Keywords: Tourist recommendation system · Collaborative-based filtering · Demographic-based filtering.

1 Introduction

Tourism in Mexico has a significant impact on the economy due to its multiplier effects in the generation of added value and employment [13]. Only in 2019, the tourism industry contributed 17.2% of gross domestic product (GDP) in Mexico [7], obtaining sixth place in the international tourism ranking [18]. In terms of economic income, tourism in Mexico represents approximately 22.5 billion dollars per year [21]. Recently, the economic impact generated by the SARS-CoV-2 coronavirus pandemic has repercussions that may extend into the medium term [2, 6, 23]. Despite this, digital technologies have allowed a reorientation of the social, cultural, and economic models related to the tourism proposals that would alleviate such impact.

Currently, many technologies allow and achieve the scope of tourism at all its levels (transport, restaurants, hotels, events, among others). However, a large

B.2. Publicación alterna al trabajo de tesis

Derivado del trabajo alterno descrito en el Apéndice A, donde se participó en el concurso para resolver los retos de identificar noticias falsas (*fake news*) y el análisis de agresividad (*aggressiveness*), se publicó el trabajo titulado "*TecNM at MEX-A3T 2020: Fake News and Aggressiveness Analysis in Mexican Spanish*" en los *Proceedings of the Iberian Languages Evaluation Forum* (IberLEF 2020).

TecNM at MEX-A3T 2020: Fake News and Aggressiveness Analysis in Mexican Spanish

Samuel Arce-Cardenas[a], Daniel Fajardo-Delgado[a] and Miguel Á. Álvarez-Carmona[b,c]

[a]*Tecnológico Nacional de México / Campus Ciudad Guzmán, Mexico.*
[b]*Centro de Investigación Científica y de Educación Superior de Ensenada, Mexico*
[c]*Consejo Nacional de Ciencia y Tecnología (CONACYT), Mexico*

Abstract
This paper describes our participation in the MEX-A3T 2020 for the tasks of identification of aggressiveness and fake news in Mexican Spanish tweets. We evaluate the combination of basic text classification techniques, including six machine learning algorithms, two methods for keyword extractions, and two preprocessing techniques. Our best run showed an F1-macro score of 0.754 for aggressiveness and 0.815 for fake news. Our preliminary results are satisfactory and competitive with other participating teams.

Keywords
Aggressiveness Identification, Fake News Classification, Natural Language Processing

1. Introduction

In today's digital culture, people spend more time on online social networks as a medium to interact, share, and collaborate with others using a style of informal communication [1]. However, these social networks are not exempt from unappropriated conducts and misbehaviors intended to cause emotional pain or to harm society through the communication process [2]. One of these destructive features of communications is the aggressiveness, a trait that involves attacking the self-concept of others [3]. The other one lies in the threat of disinformation, designed to negatively influence people and provide them an incorrect insight into different situations [4]. Both of these problems are tasks covered on the MEX-A3T 2020, a forum designed to encourage research on the analysis of social media content in Mexican Spanish [5][6].

In this work, we approach the tasks of aggressiveness and fake news posed by the MEX-A3T 2020 from a machine-learning perspective. Each of the tasks represents a binary classification problem for text content written in Mexican Spanish. The corpus for the aggressiveness task consists of 6593 tweet feeds geolocated in Mexico City. On the other hand, the corpus for the fake news task consists of 637 texts collected from January to July of 2018 from newspaper websites, media companies, and other particular websites. This work is motivated to evaluate when using basic text classification techniques is enough to provide competitive results.

Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)
EMAIL: samuel11290806@itcg.edu.mx (S. Arce-Cardenas); dfajardo@itcg.edu.mx (D. Fajardo-Delgado); malvarezc@cicese.mx (M.Á. Álvarez-Carmona)
ORCID: 0000-0002-2547-0047 (S. Arce-Cardenas); 0000-0001-8215-5927 (D. Fajardo-Delgado); 0000-0003-4421-5575 (M.Á. Álvarez-Carmona)

B.3. Producto alterno al trabajo de tesis

Se participó apoyando en la organización de la tarea REST-MEX 2021 del Iberian Languages Evaluation Forum (IberLEF 2021). Derivado de esta participación se sometió el trabajo titulado "*Overview of Rest-Mex at IberLEF 2021: Recommendation System for Text Mexican Tourism*" a la revista de la Sociedad Española para el Procesamiento de Lenguaje Natural (SEPLN).

Overview of Rest-Mex at IberLEF 2021: Recommendation System for Text Mexican Tourism

Resumen de la tarea Rest-Mex en IberLEF 2021: Sistemas de recomendación para textos turísticos mexicanos

Miguel Á. Álvarez-Carmona[1,2], Ramón Aranda[1,2], Samuel Arce-Cardenas[3], Daniel Fajardo-Delgado[3], Rafael Guerrero-Rodríguez[4], A. Pastor López-Monroy[5], Juan Martínez-Miranda[1,2], Humberto Pérez-Espinosa[1,2], Ansel Y. Rodríguez-González[1,2]
[1]Centro de Investigación Científica y de Educación Superior de Ensenada
[2]Consejo Nacional de Ciencia y Tecnología
[3]Tecnológico Nacional de México Campus Ciudad Guzmán
[4]Universidad de Guanajuato
[5]Centro de Investigación en Matemáticas
{malvarez, aranda, ansel, hperez, jmiranda}@cicese.edu.mx
r.guerrero-rodriguez@ugto.mx, pastor.lopez@cimat.mx
{daniel.fd, samuel11290806}@cdguzman.tecnm.mx

Abstract: This paper presents the framework and results from the Rest-Mex track at IberLEF 2021. This track considered two tasks: Recommendation System and Sentiment Analysis, using texts from Mexican touristic places. The Recommendation System task consists in predicting the degree of satisfaction that a tourist may have when recommending a destination of Nayarit, Mexico, based on places visited by the tourists and their opinions. On the other hand, the Sentiment Analysis task predicts the polarity of an opinion issued by a tourist who traveled to the most representative places in Guanajuato, Mexico. For both tasks, we have built new corpora considering Spanish opinions from the TripAdvisor website. This paper compares and discusses the results of the participants for both tasks.
Keywords: Rest-Mex 2021, Recommendation System, Sentiment Analysis, Mexican Tourist Text.

Resumen: Este artículo presenta los resultados de la tarea del Rest-Mex en IberLEF 2021. Este evento consideró dos sub tareas, Sistema de Recomendación y Análisis de Sentimientos, ambas utilizando textos turísticos de lugares con interés turístico en México. La tarea del Sistema de Recomendación consiste en predecir el grado de satisfacción que tendrá un turista al recomendar un destino de Nayarit, México, a partir del historial de los lugares visitados por el turista y las opiniones que se le dan a cada uno de ellos. Por otro lado, la tarea de Análisis de Sentimiento consiste en predecir la polaridad de una opinión emitida por un turista que viajó a los lugares más representativos de Guanajuato, México. Para ambas tareas, se han construido dos nuevas colecciones utilizando las opiniones en español del sitio web TripAdvisor. Este artículo compara y analiza los resultados de los participantes para ambas tareas.
Palabras clave: Rest-Mex 2021, Sistemas de recomendación, Análisis de sentimientos, Textos Turísticos Mexicanos.

1 Introduction

Tourism is a social, cultural, and economic phenomenon related to people's movement to places outside their usual place of residence for personal or business/professional reasons (Di-Bella, 2019). This activity is vital in various countries, including Mexico[1],

[1]Mexico is in the world top ten and the second Iberoamerican country related to the arrival of internatinal tourists.


The Competition Chair of the International Rest-Mex Forum
AWARD THIS CERTIFICATE TO
Samuel Arce Cárdenas
For his valuable participation as organizer of Rest-Mex 2021
Tepic, Nayarit, Mexico
July 8th, 2021
Dr. Miguel Ángel Álvarez Carmona
COMPETITION CHAIR
REST-MEX 2021
CONACYT
CIMAT
Campus Guanajuato
División de Ciencias
Económico Administrativas
Departamento de Gestión y
Dirección de Empresas