VOLUMEN 48
XXXI Verano De la Ciencia
ISSN 2395-9797
www.jovenesenlaciencia.ugto.mx

XidooTopic, del ruido semántico a los tópicos coherentes

XidooTopic: From Semantic Noise to Coherent Topics

José Adrián Rodríguez-González, Edgar Alfaro-Hernández, Jonatan Daniel Navarrete-Gómez, Rafael Guerrero-Rodríguez[2], Angel Díaz-Pacheco[1*]

[1] Departamento de Ingeniería Electrónica, División de Ingenierías, Campus Irapuato-Salamanca, Universidad de Guanajuato, México.
[2] Departamento de Gestión y Dirección de Empresas, División de Ciencias Económico Administrativas, Campus Guanajuato, Universidad de Guanajuato, México.
ja.rodriguez@ugto.mx, g.e.alfarohernandez@ugto.mx, jd.navarretegomez@ugto.mx, angel.diaz@ugto.mx

Resumen

El modelado de tópicos permite descubrir automáticamente los principales temas presentes en grandes colecciones de texto; sin embargo, su desempeño disminuye cuando se aplica a corpus con alta variabilidad léxica, como las reseñas turísticas generadas por usuarios. En este trabajo se presenta **XidooTopic**, un método de modelado de tópicos que combina reducción de dimensionalidad, detección de ruido mediante **Isolation Forest**, agrupamiento con el algoritmo de Leiden, fusión semántica de comunidades y extracción de palabras clave con **KeyBERT**. Para evaluar su desempeño se analizaron más de 37 mil reseñas en español de cinco de los principales destinos turísticos de México y se comparó con LDA y BERTopic utilizando como referencia un conjunto de tópicos obtenido por consenso de analistas humanos. Los resultados muestran que **XidooTopic** identifica una mayor proporción de los temas relevantes, genera tópicos más coherentes e interpretables y reduce tanto la fragmentación temática como la redundancia observada en los métodos de referencia.

Palabras clave: Modelado de tópicos, turismo, BERTopic, LDA, resistencia al ruido.

Introducción

El modelado de tópicos es una técnica de minería de texto que permite identificar estructuras temáticas latentes en colecciones de documentos a partir de los patrones de co-ocurrencia de sus términos (Barde & Bainwad, 2017). A diferencia de los enfoques supervisados, no requiere etiquetas previas, lo que la convierte en una herramienta especialmente adecuada para la exploración de corpus de gran escala.

Su principal ventaja radica precisamente en esa capacidad de escala: permite descubrir los temas predominantes en volúmenes de texto que resultarían inabordables mediante revisión manual, eliminando además la variabilidad y los sesgos inherentes a la interpretación humana.

No obstante, su desempeño está condicionado por la naturaleza del corpus. En colecciones de texto con alta cohesión temática, como artículos enciclopédicos o noticias, los tópicos resultantes suelen ser coherentes e interpretables. En cambio, cuando el corpus contiene ruido léxico e información subjetiva —como ocurre con las reseñas generadas por usuarios en plataformas turísticas— la calidad e interpretabilidad de los resultados se deteriora de manera considerable.

Este problema se manifiesta en dos niveles relacionados. El primero concierne a la determinación del número de tópicos: aunque existen métricas de coherencia como UMass, basada en la co-ocurrencia intra-corpus, o UCI y NPMI, fundamentadas en información puntual mutua, ninguna de ellas elimina la dependencia del criterio del analista para validar los resultados. El segundo problema, de mayor relevancia práctica, es la interpretabilidad de los tópicos resultantes: cada tópico se representa como una lista de palabras clave cuya relación semántica el analista debe inferir. Cuando dicha relación no es discernible —como en una lista que contiene términos tan dispares como limón, extraterrestre, termo y computadora— resulta imposible determinar con certeza si se trata de un tópico genuino o de ruido estadístico, dada la ausencia de conocimiento sobre el número real de temas presentes en el corpus.

VOLUMEN 48
XXXI Verano De la Ciencia
ISSN 2395-9797
www.jovenesenlaciencia.ugto.mx

De acuerdo con las revisiones sistemáticas de Álvarez-Carmona *et al*. (2022) sobre el uso de técnicas de procesamiento de lenguaje natural (PLN) en el ámbito turístico, y de Díaz-Pacheco *et al*. (2022) sobre métodos para el estudio del constructo de imagen de destino, el modelado de tópicos en este dominio estuvo dominado durante la primera mitad de la década de 2010 por el algoritmo Latent Dirichl*et al*location (LDA). Con la consolidación de las arquitecturas Transformer en el aprendizaje profundo, la segunda mitad de la década fue testigo de la adopción creciente de BERTopic, cuyas representaciones semánticas densas ofrecen ventajas sustanciales frente a los modelos probabilísticos clásicos.

Latent Dirichl*et al*location (LDA)

LDA es un modelo probabilístico generativo introducido por Blei, Ng y Jordan (2003) diseñado para descubrir estructuras temáticas latentes en grandes colecciones de texto. Su supuesto fundamental es que cada documento del corpus puede describirse como una mezcla de un número finito de tópicos, y que cada tópico, a su vez, se caracteriza por una distribución de probabilidad sobre el vocabulario (Blei *et al*., 2003).

El proceso generativo de LDA puede entenderse de la siguiente manera: para cada documento, se extrae una distribución de tópicos; para cada palabra del documento, se selecciona un tópico a partir de esa distribución; y finalmente, se elige una palabra concreta a partir de la distribución de probabilidad condicionada al tópico seleccionado. La inferencia de los parámetros del modelo se realiza mediante métodos bayesianos, como el muestreo de Gibbs o la inferencia variacional.

En LDA, un documento es visto como una combinación de diferentes tópicos: cada documento es asumido para tener un grupo de tópicos que son asignados mediante la distribución de Dirichlet, de donde toma su nombre el algoritmo. Esta distribución actúa como una prior sobre la mezcla de tópicos, lo que confiere al modelo mayor capacidad de generalización y una menor susceptibilidad al sobreajuste en comparación con enfoques previos como el Análisis Semántico Latente Probabilístico (pLSA).

Una limitación importante de LDA es que el número de tópicos debe ser especificado de antemano por el analista, lo que introduce una dependencia del criterio humano. Además, al tratar cada palabra de forma independiente (modelo bag-of-words), el algoritmo no captura relaciones semánticas entre términos contextualmente relacionados pero léxicamente distintos.

BERTopic

BERTopic es un algoritmo de modelado de tópicos presentado por Grootendorst (2022) que aborda las limitaciones de LDA al incorporar representaciones contextuales de documentos mediante modelos de lenguaje basados en la arquitectura Transformer, como BERT y sus variantes. A diferencia de LDA, BERTopic plantea el modelado de tópicos como un problema de agrupamiento (clustering) sobre embeddings de alta densidad semántica (Grootendorst, 2022).

El pipeline de BERTopic ópera en tres etapas principales. En la primera, cada documento es convertido en un vector de embeddings de alta dimensión mediante un modelo de lenguaje pre entrenado, lo que permite capturar el significado contextual de las palabras, no solo su frecuencia. En la segunda etapa, la dimensionalidad de estos embeddings es reducida mediante UMAP (Uniform Manifold Approximation and Projection), preservando la estructura semántica local. Finalmente, los documentos reducidos son agrupados mediante HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise), un algoritmo jerárquico basado en densidad que no requiere especificar el número de clusters de antemano y que tiene la capacidad de etiquetar documentos como ruido cuando no pertenecen a ningún grupo coherente.

Para la representación de cada tópico, BERTopic emplea una variante del esquema TF-IDF denominada c-TF-IDF (class-based TF-IDF), que trata a todos los documentos de un mismo cluster como si fueran un único documento de clase, identificando las palabras más discriminativas para cada tópico respecto al resto del corpus. Este enfoque genera representaciones más coherentes e interpretables que las obtenidas con LDA en dominios ruidosos como las reseñas turísticas.

No obstante, BERTopic presenta sus propias limitaciones. Al depender de HDBSCAN, tiende a generar un elevado número de micro-tópicos, especialmente en corpus grandes y heterogéneos, muchos de los cuales

VOLUMEN 48
XXXI Verano De la Ciencia
ISSN 2395-9797
www.jovenesenlaciencia.ugto.mx

son semánticamente similares entre sí y difíciles de distinguir sin una fusión posterior. Esta deficiencia fue documentada en el trabajo de Guerrero-Rodriguez *et al*. (2025), donde se identificó la proliferación de clusters con alta similitud semántica en el análisis de comentarios turísticos de la ciudad de Guanajuato.

Aplicaciones en el dominio turístico

En problemas relacionados al turismo, LDA fue empleado en el trabajo de Olmos-Martinez *et al*. (2023) para analizar los temas principales discutidos en noticias referentes a la ciudad de Cancún con un nivel moderado de precisión; sin embargo, en el trabajo de Guerrero-Rodríguez *et al*. (2023) realizado sobre el mismo corpus, los resultados fueron más claros e interpretables, permitiendo una mejor extracción de información y el descubrimiento de tópicos que LDA no fue capaz de encontrar. Por otro lado, en el trabajo de Guerrero-Rodriguez *et al*. (2025), se realizó un estudio de los comentarios turísticos de la ciudad de Guanajuato en diversos portales de Internet, donde se pone de manifiesto una de las principales deficiencias de técnicas como BERTopic al propiciar la creación de micro-tópicos y encontrar clusters semánticamente muy similares, los cuales son susceptibles de ser mezclados para mejorar su interpretación.

En respuesta a las limitaciones descritas, el presente trabajo propone XidooTopic, un método alternativo de modelado de tópicos que combina técnicas de reducción de dimensionalidad, detección de ruido, construcción de grafos de similitud y algoritmos de detección de comunidades para producir tópicos más cohesivos, separados y fácilmente interpretables. El nombre del método hace referencia a Xidoo, la denominación en lengua otomí de la ciudad de Salamanca, Guanajuato, sede del proyecto.

La organización de este documento es como se detalla a continuación: en la sección 2 se describen los conjuntos de datos utilizados en los experimentos y los elementos que componen al método propuesto; la sección 3 presenta los experimentos realizados; la sección 4 presenta la discusión; y la sección 5 las conclusiones y el trabajo futuro.

Materiales y Métodos

Con el propósito de evaluar el método propuesto, se construyó un corpus de reseñas en español sobre destinos turísticos en México, recolectadas de tres plataformas digitales de contenido generado por usuarios: TripAdvisor y Instagram. La selección de los destinos se realizó con base en criterios estadísticos oficiales: se eligieron los cinco destinos con mayor porcentaje de ocupación hotelera total según el Compendio Estadístico del Turismo en México 2024 (SECTUR, 2024). Los destinos resultantes fueron la Riviera Maya, la Riviera Nayarit, Puerto Vallarta, La Paz (Baja California Sur) y Huatulco. La distribución del corpus por destino y plataforma se detalla en la Tabla 1.

Tabla 1. *Distribución del corpus por destino turístico y plataforma digital.*

Destino	TripAdvisor	Instagram
Riviera Maya	11000	1553
Riviera Nayarit	3320	1252
Puerto Vallarta	8607	1400
La Paz	4088	353
Huatulco	3759	2423
TOTAL	30774	6981

Fuente: elaboración propia a partir de los datos recolectados.

En las siguientes secciones se detallan los elementos que conforman al método propuesto.

Isolation forest

La detección y eliminación de valores atípicos (outliers) constituye una etapa crítica en el preprocesamiento de corpus ruidosos. En el presente trabajo se emplea el algoritmo Isolation Forest (Liu *et al*., 2008) para este propósito. A diferencia de los métodos tradicionales basados en distancia o densidad, Isolation Forest adopta un enfoque basado en el aislamiento: el principio subyacente es que las anomalías son instancias escasas y

ISSN 2395-9797
www.jovenesenlaciencia.ugto.mx

estructuralmente distintas de la mayoría, por lo que resultan más fáciles de aislar mediante particiones aleatorias del espacio de características. El algoritmo construye un conjunto de árboles de aislamiento (isolation trees) en los que cada nodo divide el espacio de manera aleatoria; las instancias que requieren menos particiones para ser aisladas reciben una puntuación de anomalía más alta. Liu *et al*. (2012) demostraron que este enfoque presenta ventajas sustanciales frente a métodos como LOF (Local Outlier Factor) o SVM de una clase, entre ellas una complejidad computacional lineal respecto al tamaño del corpus y un desempeño robusto en espacios de alta dimensionalidad, condiciones ambas que caracterizan a los embeddings de texto empleados en el presente trabajo.

Algoritmo de Leiden

Para la detección de comunidades en el grafo de similitud construido sobre los embeddings, se emplea el algoritmo de Leiden (Traag *et al*., 2019). Este algoritmo constituye una mejora directa sobre el ampliamente utilizado algoritmo de Louvain (Blondel *et al*., 2008), corrigiendo una de sus deficiencias más conocidas: la posibilidad de producir comunidades internamente desconectadas, lo que compromete la coherencia estructural de los grupos resultantes. Leiden garantiza que todas las comunidades detectadas sean subgrafos conexos, lo cual es especialmente relevante en grafos de similitud semántica donde la conectividad local refleja proximidad temática entre documentos. El algoritmo opera mediante un proceso iterativo de tres fases: movimiento local de nodos, refinamiento de la partición y agregación del grafo, lo que le permite explorar el espacio de particiones de manera más eficiente que Louvain y converger hacia soluciones de mayor calidad. Traag *et al*. (2019) demostraron que Leiden alcanza valores superiores de modularidad en menos iteraciones, manteniendo al mismo tiempo una complejidad computacional casi lineal respecto al número de aristas del grafo, lo que lo hace adecuado para grafos de gran escala como los que emergen del análisis de corpus turísticos extensos.

Reducción de clusters irrelevantes y fusión semántica

Una vez obtenida la partición inicial mediante el algoritmo de Leiden, se aplica un criterio de filtrado para eliminar los clusters de tamaño reducido. Este paso está motivado por la naturaleza del propio algoritmo: al operar sobre grafos de similitud con resolución ajustable, Leiden puede generar un número elevado de micro-comunidades que, si bien son estructuralmente coherentes desde la perspectiva del grafo, carecen de suficiente respaldo estadístico para ser consideradas tópicos genuinos. En la literatura de modelado de tópicos, la eliminación de clusters pequeños es una práctica habitual para mejorar la interpretabilidad de los resultados y reducir el ruido en la representación final (Grootendorst, 2022; Allahyari *et al*., 2017). En el presente trabajo, un cluster es descartado cuando su tamaño es inferior a un umbral mínimo definido como una fracción del total de documentos del corpus, garantizando así que únicamente los tópicos con representación estadística suficiente sean considerados en las etapas posteriores del análisis.

Posteriormente, se aplica una etapa de fusión semántica sobre los clusters supervivientes con el objetivo de consolidar tópicos que, pese a haber sido detectados como comunidades distintas por el algoritmo de Leiden, presentan una alta similitud en su contenido temático. Para ello, cada cluster es representado mediante su centroide en el espacio de embeddings, calculado como la media aritmética de los vectores de sus documentos constitutivos, y normalizado a norma unitaria. La similitud entre pares de clusters se cuantifica mediante la similitud coseno entre sus centroides respectivos, una medida ampliamente adoptada en tareas de recuperación de información y agrupamiento semántico por su invarianza ante la magnitud de los vectores y su efectividad en espacios de alta dimensionalidad (Han *et al*., 2012). Aquellos pares de clusters cuya similitud coseno supera un umbral dinámico —definido a partir de la distribución empírica de similitudes inter-cluster mediante un percentil elevado, ponderado por la cohesión interna de cada cluster— son fusionados de manera iterativa, asignando todos los documentos del cluster absorbido al cluster receptor. Este proceso se repite hasta que ningún par de clusters activos supera el umbral de fusión establecido, produciendo una partición final más compacta, semánticamente homogénea y libre de redundancias temáticas.

Extracción de palabras clave mediante KeyBERT

Una vez consolidada la partición final de clusters, se procede a la extracción de las palabras clave representativas de cada tópico. Para ello se emplea KeyBERT (Grootendorst, 2020), un método de extracción de palabras clave basado en embeddings que aprovecha la capacidad de los modelos de lenguaje pre entrenados para capturar similitud semántica entre términos y documentos.

El principio de funcionamiento de KeyBERT consiste en generar embeddings tanto para el documento de entrada como para los candidatos a palabras clave, seleccionando aquellos términos cuya representación vectorial es más similar a la del documento completo. A diferencia de enfoques clásicos como TF-IDF, KeyBERT no se limita a identificar términos frecuentes, sino que es capaz de seleccionar palabras que son semánticamente centrales aunque no sean necesariamente las más frecuentes en el texto (Grootendorst, 2020). Adicionalmente, mediante el uso de Maximal Marginal Relevance (MMR) (Carbonell & Goldstein, 1998), KeyBERT permite controlar la diversidad del conjunto de palabras clave resultante, penalizando la selección de términos redundantes entre sí y favoreciendo una representación más rica y variada del contenido temático de cada cluster.

En el presente trabajo, KeyBERT no se aplica sobre la totalidad de los documentos de cada cluster, sino exclusivamente sobre el subconjunto correspondiente al 35% de documentos más cercanos al centroide del cluster en el espacio de embeddings. Esta decisión está motivada por dos razones: en primer lugar, los documentos más próximos al centroide son los más representativos del tópico y los menos contaminados por ruido periférico; en segundo lugar, restringir el análisis a este subconjunto reduce considerablemente el costo computacional del proceso, lo cual es especialmente relevante en entornos sin acceso a GPU. Para cada documento del subconjunto se extraen las diez palabras clave más relevantes, empleando como vectorizar un modelo de bolsa de palabras con eliminación de stopwords en español.

Las listas de palabras clave obtenidas para todos los documentos del subconjunto se consolidan en una única lista agregada, sobre la cual se aplica un ordenamiento por frecuencia de aparición. Con el fin de evitar la redundancia léxica introducida por variantes morfológicas de un mismo término, la lista resultante es sometida a un proceso de lematización mediante el modelo es_core_news_sm de spaCy, que reduce cada palabra a su forma canónica. Finalmente, los primeros 512 lemas del ranking son concatenados y procesados para generar una descripción compacta del contenido temático de cada cluster, la cual complementa a las palabras clave en la representación final del tópico.

Resumen del proceso

La Figura 1 presenta el diagrama de flujo completo del método XidooTopic[1]. El pipeline puede describirse como una secuencia de seis etapas principales que transforman progresivamente una colección de reseñas en bruto en un conjunto de tópicos cohesivos, interpretables y libres de redundancia semántica.

La primera etapa corresponde al preprocesamiento textual, en la que las reseñas son normalizadas, despojadas de ruido léxico y filtradas por longitud mínima, garantizando que únicamente los documentos con contenido semántico suficiente participen en las etapas posteriores.

En la segunda etapa se generan los embeddings documentales, que proyectan cada documento en un espacio vectorial denso de 384 dimensiones capturando su significado semántico. Dado que trabajar directamente en este espacio de alta dimensionalidad resulta computacionalmente costoso y perjudicial para la calidad del agrupamiento, en la tercera etapa se aplica reducción de dimensionalidad mediante UMAP, comprimiendo los embeddings a cinco dimensiones preservando la estructura semántica local del corpus.

La cuarta etapa introduce un mecanismo de detección y eliminación de ruido mediante Isolation Forest, que identifica y descarta las instancias atípicas del espacio reducido antes de que estas puedan distorsionar la estructura del grafo. Sobre los embeddings filtrados se construye en la quinta etapa un grafo de similitud k-NN utilizando distancia coseno a través de FAISS, sobre el cual se ejecuta el algoritmo de Leiden para detectar comunidades semánticamente cohesivas. Esta etapa incluye además la búsqueda de la resolución

[1] https://github.com/JoseAdrianRodriguezGonzalez/xidooTOPIC

óptima del grafo mediante una exploración sistemática que maximiza simultáneamente la cohesión intra-cluster y la separación inter-cluster.

Finalmente, la sexta etapa comprende un proceso de refinamiento de la partición en dos pasos sucesivos: primero, la eliminación de clusters de tamaño insuficiente; segundo, la fusión iterativa de clusters semánticamente redundantes mediante similitud coseno entre centroides. El resultado es un conjunto de tópicos representados por sus palabras clave más discriminativas, extraídas mediante KeyBERT sobre los documentos más cercanos a cada centroide y consolidadas mediante lematización y ordenamiento por frecuencia.

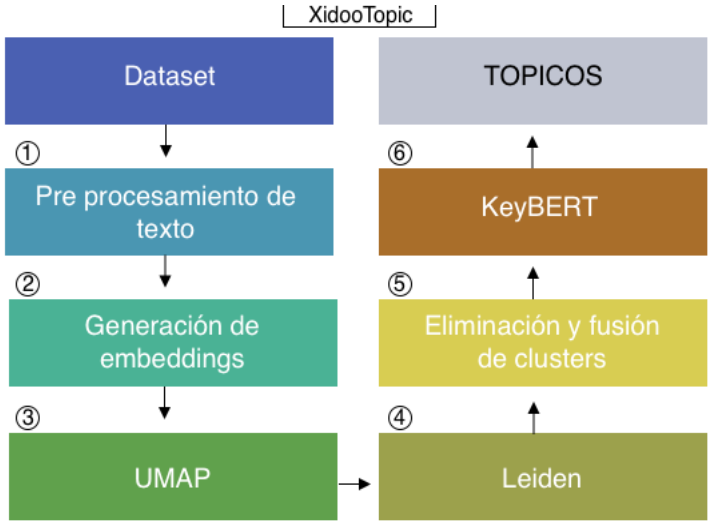


Figura 1. Diagrama de Flujo de XidooTopic.

Experimentos y resultados

La evaluación experimental fue diseñada para verificar la capacidad del método XidooTopic propuesto para identificar temas representativos a partir de corpus textuales y para comparar su rendimiento frente a dos enfoques de modelado de temas ampliamente utilizados: Latent Dirichl*et al*location (LDA) y BERTopic. Para garantizar una evaluación exhaustiva, los tres métodos se aplicaron a múltiples conjuntos de datos que presentan diferentes estructuras temáticas y niveles de complejidad. Los temas resultantes se analizaron en términos de su interpretabilidad, representatividad y correspondencia con los temas identificados por expertos humanos. Además, se construyó un conjunto de referencia de temas mediante un proceso de consenso que involucró a 20 analistas que examinaron de forma independiente cada conjunto de datos utilizando diferentes flujos de trabajo analíticos. Este consenso sirve como referencia cualitativa para evaluar en qué medida los temas identificados automáticamente capturan el contenido temático subyacente de los datos.

A continuación se detallan los tópicos identificados por los expertos humanos en los corpus de los destinos investigados:

Riviera Maya

- Cenotes, cuevas y ríos subterráneos: recorridos guiados por Río Secreto, Chaak Tun y Dos Ojos, con snorkel, cavernas y formaciones naturales.
- Snorkel con tortugas en Akumal: experiencia específica de avistamiento de tortugas marinas en la bahía, usualmente con guías y equipo rentado.
- Experiencias en hoteles y resorts: reseñas sobre habitaciones, servicio, comida, amenidades de playa y actividades familiares en complejos turísticos.

- Parques temáticos y atracciones turísticas: visitas a Xcaret y otros parques, espectáculos, recorridos naturales y experiencias organizadas.
- Interacción con fauna y zoológicos: experiencias tipo CrocoCun, contacto con cocodrilos, monos, serpientes y otros animales en recorridos familiares.
- Ruinas mayas y turismo urbano: visita a Tulum, Playa del Carmen, zonas comerciales, transporte, compras y restaurantes.

Riviera Nayarit

- Playas de Sayulita y Nuevo Vallarta: caminatas costeras, arena, oleaje, atardeceres y descanso frente al mar.
- Parques acuáticos y experiencias con delfines: nado con delfines, interacciones con lobos marinos, toboganes, espectáculos familiares y actividades organizadas.
- Santuario de cocodrilos y conservación de fauna: recorridos educativos con biólogos, observación de cocodrilos y experiencias de conservación.
- Restaurantes frente al mar y beach clubs: comida en la playa, camastros, sombrillas y experiencias relajadas junto al mar.
- Mercados locales y puerto deportivo: gastronomía, artesanías, productos regionales, música en vivo y actividad comercial cercana a la marina.
- Ambiente familiar y recreación social: paseos con familia o amigos, convivencia, descanso y ocio general en espacios turísticos.

Puerto Vallarta

- Malecón, arte urbano y paseo costero: caminatas frente al mar, esculturas, estatuas, arte callejero, atardeceres y ambiente social.
- Playas y deportes acuáticos: Playa de los Muertos, actividades en la arena, parasailing, jetskis y vendedores ambulantes.
- Vida nocturna, gastronomía y compras: bares, restaurantes, tiendas, vendedores, vida urbana turística y convivencia en la zona del Malecón.
- Actividad portuaria, marina y tours en barco: puerto deportivo, embarcaciones, cruceros, paseos por la bahía y excursiones marítimas.
- Tours de aventura y degustación de tequila: tirolesas, recorridos guiados, fincas de agave, tequila, mezcal y raicilla.
- Catedral e iglesias coloniales: arquitectura religiosa, misas, iglesias icónicas, historia local y centro histórico.

La Paz

- Isla Espíritu Santo y vida marina: snorkel con lobos marinos, avistamiento de ballenas y tortugas, paseos en bote y paisajes naturales.
- Malecón de La Paz, arte urbano y esculturas: caminatas, vistas al mar, esculturas, ambiente familiar, atardeceres y paseo urbano.
- Playa Balandra: aguas cristalinas someras, el hongo de piedra, kayak, caminatas en la bahía y paisaje natural protegido.
- Tours guiados en bote: excursiones de día completo, capitanes locales, almuerzo incluido, servicio personalizado y recorridos por playas o islas.
- Movilidad recreativa y paseo en bicicleta: recorridos en bicicleta, paseos tranquilos y actividades recreativas al aire libre.
- Catedral y centro histórico: iglesias coloniales, arquitectura religiosa, plazas, historia local y puntos centrales de la ciudad.

ISSN 2395-9797
www.jovenesenlaciencia.ugto.mx

Huatulco

- Snorkel en bahías y arrecifes: aguas claras, arrecifes de coral, peces tropicales, equipo de snorkel y actividades acuáticas.
- Playas, paisaje costero y arena negra: caminatas por la playa, vistas naturales, arena blanca o negra, paisajes volcánicos, descanso y relajación frente al mar.
- Restaurantes frente al mar y beach clubs: comida en la playa, camastros, sombrillas, restaurantes costeros y ambiente relajado.
- Puerto deportivo, marina y cruceros: clubes de playa, restaurantes frente al muelle, embarcaciones, tours en barco y llegada de cruceros.
- Experiencias fuera de la playa: baños de barro, tours de café, ríos, comunidades locales y encuentros río-océano.

Comparación de modelado de tópicos en las diferentes técnicas consideradas

Las Tablas 2 a 6 presentan las palabras clave identificadas por los tres enfoques evaluados —XidooTopic (el método propuesto), Latent Dirichl*et al*location (LDA) y BERTopic— para cada conjunto de datos analizado. Durante los experimentos, BERTopic mostró una tendencia a generar un gran número de microtemas altamente específicos, lo que dio lugar a un extenso conjunto de palabras clave asociadas. Por lo tanto, para facilitar la comparación entre métodos y evitar aumentar innecesariamente la extensión del manuscrito, solo se reporta un subconjunto limitado de los resultados más representativos de BERTopic. Esta comparación permite evaluar la capacidad de cada método para identificar los temas principales presentes en los datos y su nivel de concordancia con las interpretaciones derivadas de los expertos.

Tabla 2. *Tópicos identificados por LDA, BERTopic y XidooTopic para el corpus de Riviera Maya.*

LDA	BERTopic	XidooTopic
cueva, guía, agua, recorrido, foto, tomar, experiencia, increíble	excelente, realmente, como, fotógrafo, recorrido, una, precio, que	increíble, guía, personal, lugar, familia, gran, servicio, excelente
playa, tortuga, guía, snorkel, área, agua, pagar, nadar	cenote, cenotes, pesos, formaciones, realmente, definitivamente, segundo, como	maya, pesos, carmen, ruina, tulum, taxi, resort, hotel
habitación, hotel, buen, playa, comida, servicio, parque, excelente	snorkeling, snorkel, tortugas, realmente, excelente, buceo, hermoso, viaje	playa, tortuga, akumal, área, ver, bahía, tours, peso
tienda, lugar, playa, tomar, agua, caminar, gente, foto	animales, zoológico, animal, interactivo, alimentación, interacción, cuidados, cuidado	guir, increíble, río, rio, experiencia, caminar, ver, tour
animal, personal, guía, zoológico, hotel, excelente, cocodrilo, experiencia	caverna, cavernas, formaciones, realmente, proporcionar, por, absolutamente, como	foto, hotel, guir, traje, autobús, snorkel, increíble, conductor

ISSN 2395-9797
www.jovenesenlaciencia.ugto.mx

Tabla 3. *Tópicos identificados por LDA, BERTopic y XidooTopic para el corpus de Riviera Nayarit.*

LDA	BERTopic	XidooTopic
playa, agua, caminar, arena, pequeño, hermoso, ola, lugar	playa, playas, excelente, realmente, buena, bahía, pequeños, como	caminar, gente, agua, arena, sayulitar, hermoso, gran, lugar
delfín, foto, parque, experiencia, niño, tobogán, playa, acuático	santuario, taxista, cuidado, taxi, pesos, excelente, dinero, trabajo	experiencia, niño, increíble, personal, tobogán, parcar, divertido, tiempo
delfín, mercado, experiencia, excelente, tiempo, comida, increíble, animal	mercado, delicioso, calidad, mercados, puerto, pesos, deportivo, deliciosa	arena, bahía, caminar, resort, restaurante, boogie, largo, hotel
cocodrilo, animal, lugar, recorrido, playa, santuario, tomar, bien	delfines, definitivamente, interacción, cuidado, atracciones, descubrimiento, actividades, realmente	delfín, león, familia, tobogán, experiencia, increíble, niño, hotel
delfín, parque, playa, comida, acuático, agua, tobogán, nadar	leones, decepcionados, excelente, descubrimiento, especialmente, interacción, definitivamente, acuático	cocodrilo, santuario, taxi, guir, niño, pesos, lugar, biólogo

Tabla 4. *Tópicos identificados por LDA, BERTopic y XidooTopic para el corpus de Puerto Vallarta.*

LDA	BERTopic	XidooTopic
restaurante, lugar, playa, tienda, pasear, malecón, excelente, gente	excelente, realmente, buena, como, hermoso, ciudad, puerto, por	puerto, caminar, gente, lugar, ver, arte, agradable, vallarta
pasear, tienda, caminar, malecón, restaurante, mucho, marítimo, puerto	statues, hotel, visiting, beaches, places, fun, sculptures, beach	restaurantes, puerto, playa, vallarta, artista, caminar, tienda, océano
lugar, restaurante, gente, comida, tomar, puerto, hermoso, mucho	ziplines, zipline, guides, guide, zip, lines, experience, line	gente, muerto, restaurante, lugar, arena, caminar, vendedor, gran
malecón, vallarto, pasear, restaurante, puerto, tienda, caminar, vendedor		bar, tienda, lugar, paseo, bahía, vistas, mucho, caminar
malecón, lugar, pasear, gente, tienda, caminar, mucho, recorrido		hermoso, dentro, visitar, ver, pena, ciudad, historia, malecón

Tabla 5. *Tópicos identificados por LDA, BERTopic y XidooTopic para el corpus de La Paz.*

LDA	BERTopic	XidooTopic
playa, lugar, león, agua, islar, viaje, hermoso, nadar	isla, islas, coral, leones, como, pena, bahía, realmente	león, snorkel, playa, increíble, espíritu, paz, barco, viaje
playa, agua, lugar, caminar, malecón, mucho, hermoso, islar	paz, frente, hermoso, caminando, realmente, pasarela, largo, actividad	restaurante, paz, caminar, lugar, largo, camino, gente,

		noche
playa, agua, pena, valer, caminar, lugar, pequeño, hermoso	restaurantes, buenos, restaurante, hermoso, heladerías, excelente, cafés, realmente	balandra, paz, arenar, caminar, lugar, agua, ver, estacionamiento
islar, león, agua, playa, snorkel, hacer, viaje, barco	playas, playa, hermosa, hermoso, hermosas, estacionamiento, cuidado, agua	pena, viaje, agua, hermoso, día, gran, experiencia, gente
catedral, iglesia, bonito, hermoso, visitar, historia, plaza, ciudad	verde, isla, colores, diversidad, coloridos, hermoso, como, leones	playa, colores, hermoso, isla, espíritu, increíble, viaje, barco

Tabla 6. *Tópicos identificados por LDA, BERTopic y XidooTopic para el corpus de Huatulco.*

LDA	BERTopic	XidooTopic
snorkel, restaurante, lugar, taxi, excelente, hacer, gente, bahía	snorkel, snorkeling, restaurante, restaurantes, océano, largo, hoteles, realmente	lugar , arena, agua, bahía, gente, gran, nadar, hermoso
agua, restaurante, hermoso, bahía, lugar, buen, snorkel, nadar	caminata, camino, playa, caminar, realmente, playas, excelente, cerveza	taxi, restaurantes, snorkel, peso, restaurante, viaje, excelente, día
restaurante, agua, bahía, pequeño, lugar, buen, barco, snorkel	bahía, bahías, cuidado, buena, marido, actividades, encantó, amado	restaurantes, playa, snorkel, resort, restaurante, lugar, océano, cruz
hotel, restaurante, área, nadar, snorkel, comida, buen, taxi	coral, corales, snorkeling, snorkel, escuelas, bahía, hermosos, muerto	playa, restaurantes , hacer, snorkeling, excelente, área, taxi, lugar
restaurante, increíble, agua, hermoso, barro, hacer, comida, pequeño	deportivo, puerto, clubes, restaurantes, restaurante, excelente, playa, barco	playa, restaurantes, snorkel, taxi, huatulco, pesos, arena, áreo

Con base en los resultados presentados en las Tablas 2–6 y en los tópicos de referencia definidos por los analistas humanos, se observa que **XidooTopic presenta la mayor cobertura temática de los tres métodos evaluados**. A diferencia de LDA, que en varios casos fusiona distintos temas en un mismo tópico debido a la naturaleza probabilística de la distribución de palabras, y de BERTopic, que tiende a fragmentar la información en numerosos microtópicos altamente específicos, XidooTopic logra recuperar de forma consistente la mayoría de los temas considerados relevantes por los evaluadores humanos. En Riviera Maya, por ejemplo, identifica claramente los tópicos relacionados con cenotes y ríos subterráneos, snorkel con tortugas en Akumal, hoteles y resorts, ruinas mayas de Tulum y Playa del Carmen, además de elementos asociados con las experiencias guiadas; únicamente el zoológico CrocoCun aparece parcialmente integrado dentro de un tópico más amplio. De forma similar, en Riviera Nayarit recupera los cuatro temas definidos por los expertos (playas, parques acuáticos con delfines, santuario de cocodrilos y mercados locales), mientras que en Puerto Vallarta identifica adecuadamente el malecón, las playas, la oferta gastronómica y comercial, así como el centro histórico, aunque el tópico relacionado con las actividades de aventura y degustación de tequila aparece representado de manera menos explícita.

La misma tendencia se mantiene en los demás destinos analizados. En La Paz, XidooTopic recupera los principales temas asociados con la Isla Espíritu Santo, Playa Balandra, el Malecón y el centro urbano, integrando además elementos de las excursiones en barco dentro de los tópicos relacionados con la isla, lo

VOLUMEN 48
XXXI Verano De la Ciencia
ISSN 2395-9797
www.jovenesenlaciencia.ugto.mx

que refleja una representación más cercana a la forma en que los turistas describen sus experiencias. En Huatulco, el método identifica con claridad los temas de snorkel, playas, marina y servicios turísticos, aunque las experiencias fuera de la playa (como los baños de barro o los recorridos cafeteros) aparecen parcialmente absorbidas por tópicos de mayor cobertura. En conjunto, estos resultados indican que XidooTopic no solo mantiene una alta correspondencia con el consenso alcanzado por los analistas humanos, sino que también ofrece un equilibrio más adecuado entre **cobertura temática e interpretabilidad**. Mientras que BERTopic suele incrementar la granularidad generando múltiples subtemas difíciles de sintetizar, y LDA frecuentemente combina temas conceptualmente distintos en un mismo tópico, XidooTopic produce un conjunto de temas suficientemente específico para distinguir las principales experiencias turísticas, pero lo bastante general para facilitar su interpretación por parte de analistas humanos. Esta capacidad de aproximarse al consenso humano sugiere que el método propuesto constituye una alternativa más adecuada para aplicaciones de minería de opiniones donde la exhaustividad y la facilidad de interpretación son factores igualmente importantes.

Análisis de robustez frente a la contaminación entre dominios

Para evaluar la robustez de XidooTopic ante la presencia de documentos semánticamente no relacionados, se realizó un experimento controlado de contaminación utilizando el subconjunto de Riviera Maya del conjunto de datos TripAdvisor (11.000 reseñas; Tabla 1). Como corpus externo, se emplearon 2.889 reseñas de productos electrónicos de Amazon España seleccionadas aleatoriamente (Keung *et al*., 2020). Este conjunto de datos se eligió intencionadamente porque comparte el mismo formato de reseña —textos breves, en primera persona e informales—, pero representa un dominio semántico completamente diferente.

El corpus contaminante se inyectó de forma incremental en el corpus turístico en trece niveles de contaminación (1%, 2%, 3%, 5%, 7%, 10%, 11%, 12%, 13%, 14%, 15%, 20% y 25% del tamaño del corpus original). Tras identificar la región de transición, se evaluaron cuatro niveles intermedios adicionales (16–19%) para determinar con mayor precisión el umbral de contaminación.

Se estableció un criterio de dos etapas para determinar si el tema contaminante se había identificado con éxito. La aparición de cualquier sustantivo específico del dominio (por ejemplo, batería, vendedor, Amazon, funda, garantía) se consideró evidencia suficiente de contaminación. Por el contrario, los términos genéricos relacionados con quejas que también pueden aparecer en reseñas turísticas (por ejemplo, calidad, precio, funcionamiento, comprar) solo se consideraron evidencia válida cuando al menos dos de ellos aparecían simultáneamente dentro del mismo tema.

Como se muestra en la Tabla 7, XidooTopic exhibió una transición bien definida entre el 15% y el 16% de contaminación. Por debajo de este umbral, ningún tema extraído pudo identificarse como perteneciente al dominio externo. Sin embargo, a partir del 16%, el tema contaminante se detectó de manera consistente. Para verificar que esta transición no era un artefacto de una única ejecución, se realizó un procedimiento de bootstrap generando diez submuestras aleatorias que contenían el 80% de los documentos para los dos niveles límite (Tabla 8). Las tasas de detección aumentaron del 50% con un 16% de contaminación al 90% con un 17%, lo que confirma la existencia de una región de transición estable, más que una variación aleatoria.

Tabla 7. *Detección del tópico contaminante por nivel de contaminación con XidooTopic. Solo se presentan niveles de interés para mejorar la legibilidad.*

Nivel	Existe el cluster contaminado?	Palabras Clave	Marcadores débiles (calidad/precio/comprar/funcionar)	Detectado?
14	SI	bien, **calidad**, si, perfecto, bastante, llegar, dos, pega, cumplir, esperar	1	NO
15	SI	bien, **calidad**, si, bastante, esperar, dos, solo, llegar, perfecto, hacer	1	NO

16	SI	bien, **calidad**, si, bastante, dos, **funcionar**, llegar, cumplir, esperar, recomiendo	2	SI
17	SI	**producto**, bien, **calidad**, si, buen, dos, bastante, **funcionar**, recomiendo, buena	2 + Sustantivo Fuerte	SI
18	SI	bien, si, calidad, bastante, dos, funcionar, esperar, solo, buen, hacer	2	SI
19	SI	bien, si, **calidad**, dos, bastante, **funcionar**, hacer, recomiendo, llegar, solo	2	SI
20	SI	bien, **calidad**, si, dos, bastante, buen, **precio**, recomiendo, regalo, solo	2	SI
25	SI	bien, si, **calidad**, dos, poner, bastante, solo, **funcionar**, perfecto, función	2	SI

Tabla 8.*Tasa de detección en réplicas bootstrap (10 submuestras al 80%) en los niveles frontera.*

Nivel de Contaminación	Replicas Válidas	Detecciones	Tasa de Detección
16%	10	5	50%
17%	10	9	90%

Posteriormente, el mismo experimento se repitió utilizando LDA y BERTopic en condiciones idénticas (Figura 2). A diferencia de XidooTopic, LDA no mostró un umbral de detección consistente. Se produjeron detenciones aisladas con un 5% y un 13% de contaminación, sin ninguna relación monótona con el nivel real de contaminación. Este comportamiento puede atribuirse a la optimización automática del número de temas mediante la puntuación de coherencia UMass, que favoreció sistemáticamente soluciones de granularidad gruesa que comprenden sólo de cinco a nueve temas, diluyendo así la señal léxica de los documentos contaminantes.

VOLUMEN 48
XXXI Verano De la Ciencia
ISSN 2395-9797
www.jovenesenlaciencia.ugto.mx

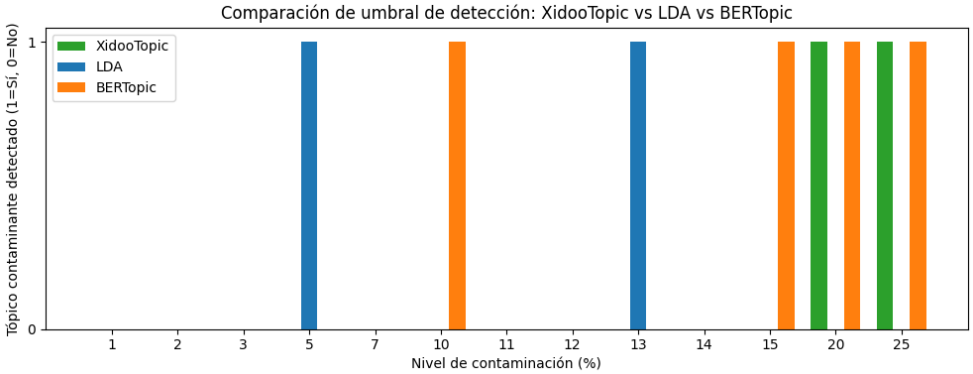


Figura 2. Comparación del umbral de detección entre XidooTopic, LDA y BERTopic.

BERTopic tampoco logró proporcionar un comportamiento de detección estable. Su rendimiento resultó ser muy sensible al parámetro tamaño mínimo del clúster de HDBSCAN. Utilizando el valor comúnmente adoptado del 5% del tamaño del corpus, BERTopic se redujo a sólo dos temas en casi todos los niveles de contaminación. Al reducir el parámetro al 1%, se obtuvieron entre dos y diecinueve temas, dependiendo del nivel de contaminación, sin una correspondencia clara con la cantidad de ruido inyectado. Este comportamiento sugiere que cambios relativamente pequeños en la composición del corpus afectan sustancialmente la estructura temática resultante, lo que indica una menor estabilidad estructural que la alcanzada por XidooTopic.

Las métricas de calidad temática respaldan aún más estas observaciones (Figura 3). XidooTopic alcanzó sistemáticamente coeficientes de silueta sustancialmente más altos que BERTopic (0,29–0,42 frente a 0,04–0,07), lo que indica agrupaciones de temas más compactas y mejor separadas en el espacio de incrustaciones (embeddings). Aunque BERTopic y LDA obtuvieron puntuaciones de coherencia textual más altas (C_v, NPMI y UMass) en varios niveles de contaminación, este resultado es esperable, ya que ambos métodos generaron considerablemente menos temas y más amplios. Tales representaciones de granularidad gruesa tienden a aumentar artificialmente las métricas de coherencia basadas en co-ocurrencias, al tiempo que proporcionan una menor discriminación semántica entre temas distintos.

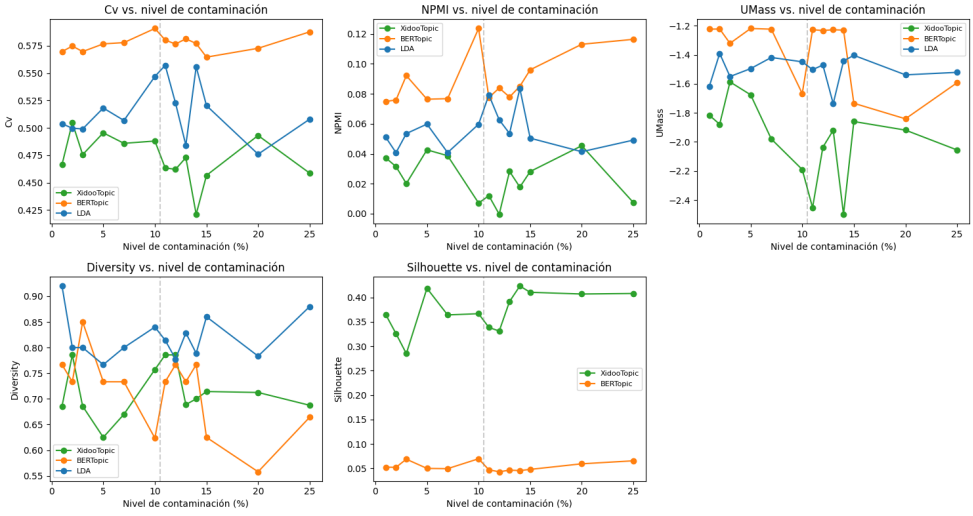


Figura 3. Evolución de las métricas de calidad de los tópicos (C_v, NPMI, UMass, Diversity, Silhouette) en función del nivel de contaminación.

Selección del LLM y análisis de interpretabilidad basado en la percepción de evaluadores humanos

Como se mencionó en la sección anterior, el objetivo de XidooTopic no consiste únicamente en generar tópicos semánticamente coherentes, sino también en facilitar su interpretación por parte de analistas humanos. Para ello, el método propuesto combina su estrategia de extracción de tópicos —caracterizada por su robustez frente al ruido documental y por la fusión de tópicos con alta similitud semántica— con el uso de modelos ligeros de lenguaje de gran escala (Large Language Models, LLMs), encargados de generar descripciones más claras y representativas para cada tópico.

Con el propósito de seleccionar el LLM más adecuado, los tópicos generados por XidooTopic fueron utilizados como entrada para los siguientes modelos: **Aitana-s2b-c0dc7-f16, ChatGPT-5-q8_0, Claude3.5-4b_Instruct, Gemini-Nano-Gemmafied.Q8_0, ChatGPT-4o-distil-Llama-3.1-8B, Meta-Llama-3.1-8B-Instruct-Q4_K_M, Phi-3-mini-4k-instruct-q4** y **Mistral-7B-v0.1.Q4_K_M**.

La evaluación se llevó a cabo utilizando el conjunto de reseñas de **TripAdvisor**, ya que corresponde al corpus con mayor número de documentos y mayor riqueza léxica, constituyendo un escenario más exigente para evaluar la interpretabilidad de los tópicos generados. Cuatro evaluadores humanos analizaron de forma independiente las descripciones producidas por cada LLM y las calificaron mediante una escala ordinal de 1 a 10, donde **1** representó el nivel más bajo de interpretabilidad y **10** el más alto.

Debido a que la evaluación involucró más de dos evaluadores y empleó una escala ordinal, la concordancia entre ellos se determinó mediante el **Alfa de Krippendorff**, una de las medidas más apropiadas para este tipo de datos. El coeficiente obtenido fue **$\alpha = 0.913$**, lo que indica un nivel de acuerdo excelente entre los evaluadores y proporciona evidencia de la alta confiabilidad de la evaluación cualitativa realizada.

De acuerdo con las calificaciones obtenidas, **Mistral-7B** generó las descripciones de tópicos más interpretables y con mayor correspondencia respecto a la línea base definida manualmente por los analistas al inicio de esta sección, alcanzando la mayor puntuación promedio (**7.4/10**). En consecuencia, este modelo fue seleccionado como el LLM empleado en el resto de los experimentos.

Con el fin de realizar una comparación justa entre enfoques, el mismo procedimiento de evaluación se aplicó a los resultados obtenidos mediante **BERTopic** y **LDA**, utilizando igualmente **Mistral-7B** para la generación de las descripciones de los tópicos (ver Tabla 9). En ambos casos, la concordancia entre evaluadores fue aceptable, obteniéndose valores del Alfa de Krippendorff de **0.691** para BERTopic y **0.783** para LDA. Respecto a la evaluación de interpretabilidad, la combinación **BERTopic–Mistral-7B** obtuvo una calificación promedio de **6.4/10**, mientras que **LDA–Mistral-7B** alcanzó **5.2/10**. En contraste, **XidooTopic–Mistral-7B** obtuvo la mayor calificación promedio (**7.4/10**), lo que sugiere que los tópicos generados por el método propuesto fueron percibidos como más comprensibles y semánticamente más cercanos a la referencia establecida por los expertos.

Es importante señalar que la interpretabilidad constituye una propiedad inherentemente subjetiva y dependiente del juicio humano. Por esta razón, los resultados obtenidos no deben considerarse como evidencia concluyente de la superioridad del método propuesto. Sin embargo, sí aportan evidencia empírica de que XidooTopic tiende a generar representaciones temáticas que son percibidas como más claras, consistentes e interpretables por evaluadores independientes, complementando así los resultados cuantitativos presentados en las secciones anteriores.

Tabla 9. *Promedios de calificaciones obtenidos por BERTopic, LDA y XidooTopic en el análisis de interpretabilidad.*

Método	LLM	Alfa de Krippendorff	Calificación promedio
BERTopic	Mistral-7B	0.691	6.4
LDA	Mistral-7B	0.783	5.2
XidooTopic	**Mistral-7B**	**0.913**	**7.4**

Conclusiones

Los experimentos realizados demuestran que XidooTopic constituye una alternativa eficaz para el modelado de tópicos en corpus turísticos caracterizados por un elevado nivel de ruido semántico y heterogeneidad léxica. La integración de técnicas de detección de anomalías, agrupamiento basado en grafos, fusión semántica y extracción contextual de palabras clave permitió obtener una representación temática más compacta y coherente que la producida por los métodos comparativos.

La comparación con LDA y BERTopic evidenció diferencias claras en el comportamiento de los tres enfoques. Mientras que LDA tendió a mezclar experiencias distintas dentro de un mismo tópico debido a su naturaleza probabilística, BERTopic produjo un número elevado de microtópicos que, aunque semánticamente consistentes, dificultan la interpretación global del corpus. En contraste, XidooTopic alcanzó un mejor equilibrio entre granularidad e interpretabilidad, recuperando la mayoría de los temas identificados por los analistas humanos sin fragmentar excesivamente la información.

La evaluación sobre los cinco destinos turísticos mostró que el método propuesto presenta una alta cobertura temática respecto al consenso de expertos, identificando de forma consistente las principales experiencias descritas por los turistas en las diferentes regiones analizadas. Esta capacidad sugiere que XidooTopic puede reducir la carga de interpretación manual y facilitar el análisis exploratorio de grandes volúmenes de opiniones generadas por usuarios.

Como trabajo futuro se plantea extender la evaluación a corpus de otros dominios y lenguajes, incorporar métricas cuantitativas adicionales de estabilidad y calidad de los agrupamientos, y explorar el uso de modelos de lenguaje de gran tamaño para generar descripciones automáticas de los tópicos detectados, incrementando aún más su utilidad para analistas e investigadores.

Referencias

Allahyari, M., *et al.* (2017). A brief survey of text mining: Classification, clustering and extraction techniques. arXiv preprint arXiv:1707.02919. https://arxiv.org/abs/1707.02919

Álvarez-Carmona, M. Á., Aranda, R., Rodríguez-Gonzalez, A. Y., Fajardo-Delgado, D., Sánchez, M. G., Pérez-Espinosa, H., Martínez-Miranda, J., Guerrero-Rodríguez, R., Bustio-Martínez, L., & Díaz-Pacheco, Á. (2022). Natural language processing applied to tourism research: A systematic review and future research directions. *Journal of King Saud University - Computer and Information Sciences,* 34(10), 10125–10144. https://doi.org/10.1016/j.jksuci.2022.10.010

Barde, B. V., & Bainwad, A. M. (2017). An overview of topic modeling methods and tools. 2017 International Conference on Intelligent Computing and Control Systems (ICICCS). Presented at the 2017 International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai. doi:10.1109/iccons.2017.8250563

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichl*et al*location. *Journal of machine Learning research*, 3(Jan), 993-1022.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Carbonell, J., & Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. En Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 335–336). ACM. https://doi.org/10.1145/290941.291025

Diaz-Pacheco, A., Álvarez-Carmona, M. Á., Guerrero-Rodríguez, R., Chávez, L. A. C., Rodríguez-González, A. Y., Ramírez-Silva, J. P., & Aranda, R. (2022). Artificial intelligence methods to support the research of destination image in tourism. A systematic review. *Journal of Experimental & Theoretical Artificial Intelligence*, 36(7), 1415–1445. https://doi.org/10.1080/0952813x.2022.2153276

Grootendorst, M. (2020). KeyBERT: Minimal keyword extraction with BERT. Zenodo. https://doi.org/10.5281/zenodo.4461265

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794. https://arxiv.org/abs/2203.05794

Guerrero-Rodríguez, R., Álvarez-Carmona, M. A., Aranda, R., & Díaz-Pacheco, Á. (2025). AI-driven analysis of hotel digital reputation: insights from a Mexican destination. *International Journal of Tourism Cities*, 12(1), 19–36. https://doi.org/10.1080/20565607.2025.2600691

Guerrero-Rodríguez, R., Álvarez-Carmona, M. Á., Aranda, R., & Díaz-Pacheco, Á. (2023). Big data analytics of online news to explore destination image using a comprehensive deep-learning approach: a case from Mexico. *Information Technology & Tourism,* 26(1), 147–182. https://doi.org/10.1007/s40558-023-00278-5

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann. https://doi.org/10.1016/C2009-0-61819-5

Keung, P., Lu, Y., Szarvas, G., & Smith, N. A. (2020). The Multilingual Amazon Reviews Corpus. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation Forest. En Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008) (pp. 413–422). IEEE. https://doi.org/10.1109/ICDM.2008.17

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2012). Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data (TKDD),* 6(1), 1–39. https://doi.org/10.1145/2133360.2133363

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.

Olmos-Martínez, E., Álvarez-Carmona, M. Á., Aranda, R., & Díaz-Pacheco, A. (2023). What does the media tell us about a destination? The Cancun case, seen from the USA, Canada, and Mexico. *International Journal of Tourism Cities*, 10(2), 639–661. https://doi.org/10.1108/ijtc-09-2022-0223

SECTUR. (1 C.E., January 10). Datatur. Compendio Estadístico Del Turismo En México. Retrieved March 26, 2026, https://datatur.sectur.gob.mx/SitePages/compendioestadistico.aspx

Traag, V. A., Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports,* 9(1), 5233. https://doi.org/10.1038/s41598-019-41695-z