UNIVERSIDAD NACIONAL DEL ALTIPLANO

ESCUELA DE POSGRADO

DOCTORADO EN ESTADÍSTICA APLICADA

TESIS

INFERENCIA LATENTE Y LA MODELIZACIÓN DE DIRICHLET PARA LA CARACTERIZACIÓN DE REVISIONES DE TESIS

PRESENTADA POR:

FRED TORRES CRUZ

PARA OPTAR EL GRADO ACADÉMICO DE:

DOCTORIS SCIENTIAE EN ESTADÍSTICA APLICADA

PUNO, PERÚ

2023

Reporte de similitud

NOMBRE DEL TRABAJO

INFERENCIA LATENTE Y MODELIZACIÓN DE DIRICHLET PARA LA CARACTERIZACIÓN DE REVISIONES DE TESIS

AUTOR

Fred Torres-Cruz

RECUENTO DE PALABRAS

19103 Words

RECUENTO DE CARACTERES

113952 Characters

RECUENTO DE PÁGINAS

71 Pages

TAMAÑO DEL ARCHIVO

1.2MB

FECHA DE ENTREGA

Sep 27, 2023 9:36 AM GMT-5

FECHA DEL INFORME

Sep 27, 2023 9:37 AM GMT-5

4% de similitud general

El total combinado de todas las coincidencias, incluidas las fuentes superpuestas, para cada base d

- 3% Base de datos de Internet
- Base de datos de Crossref
- 3% Base de datos de trabajos entregados
- 1% Base de datos de publicaciones
- Base de datos de contenido publicado de Crossr

Excluir del Reporte de Similitud

- *Material bibliográfico*
- Material citado
- Material citado
- Coincidencia baja (menos de 10 palabras)

José P. Tito Lipa
Ing. Estadístico e Informático D.Sc.
CIP. 159645

UNIVERSIDAD NACIONAL DEL ALTIPLANO - PUNO - PERÚ
ESCUELA DE POSGRADO
COORDINACIÓN DE INVESTIGACIÓN
ESTADISTICO E INFORMÁTICO
CIP. 116625

UNIVERSIDAD NACIONAL DEL ALTIPLANO

ESCUELA DE POSGRADO

DOCTORADO EN ESTADÍSTICA APLICADA

TESIS

INFERENCIA LATENTE Y LA MODELIZACIÓN DE DIRICHLET PARA LA CARACTERIZACIÓN DE REVISIONES DE TESIS

PRESENTADA POR:

FRED TORRES CRUZ

PARA OPTAR EL GRADO ACADÉMICO DE:

DOCTORIS SCIENTIAE EN ESTADÍSTICA APLICADA

APROBADA POR EL JURADO SIGUIENTE:

PRESIDENTE

.....
Dr. LEONEL COYLA IDME

PRIMER MIEMBRO

.....
Dr. ADOLFO CARLOS JIMENEZ CHURA

SEGUNDO MIEMBRO

.....
Dr. FIDEL ERNESTO TICONA YANQUI

ASESOR DE TESIS

.....
D.Sc. JOSE PANFILO TITO LIPA

Puno, 06 de Octubre de 2023

ÁREA: Inteligencia Artificial.
TEMA: Recuperación de Información Inferencia Latente y Modelización Direchlet para Revisiones Caracterización de Revisiones de Tesis.
LÍNEA: Base de Datos.

DEDICATORIA

A mis padres y mi hermano, cuyo amor incondicional y apoyo constante han sido mi faro en la tempestad.

A los profesores del Doctorado en Estadística Aplicada, por su invaluable orientación.

AGRADECIMIENTOS

A los docentes y compañeros del Doctorado en Estadística Aplicada impartieron sus conocimientos y experiencia.

A los miembros del jurado quienes han contribuido con la culminación de este estudio de manera satisfactoria.

Al Vicerrectorado de Investigación por contribuir con la información en la ejecución del estudio presentado.

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

El objetivo principal de esta investigación consistió en la aplicación de la técnica de modelización Latent Dirichlet Allocation (LDA) como un enfoque de inferencia latente para la clasificación temática de comentarios realizados por revisores de tesis. Se buscó evaluar exhaustivamente la capacidad del modelo LDA para identificar y agrupar los comentarios en diversas categorías temáticas, con el propósito de obtener una comprensión y organización más precisa de la información contenida en las revisiones de tesis. Para lograr este objetivo, se recopiló un corpus de datos compuesto por comentarios de revisores de tesis registrados en la Plataforma de Investigación Integrada a la Labor Académica con Responsabilidad (PILAR) de la prestigiosa Universidad Nacional del Altiplano de Puno (UNAP). Posteriormente, se aplicó el modelo LDA para generar un sistema de clasificación que permitiera asignar de manera efectiva las revisiones de tesis a categorías temáticas establecidas de acuerdo a una taxonomía previamente definida. Finalmente, se llevó a cabo una exhaustiva evaluación de la eficacia y precisión del modelo desarrollado, con el fin de determinar su capacidad para clasificar de manera óptima los comentarios. Los resultados obtenidos demuestran el potencial de la técnica LDA como herramienta efectiva en la clasificación temática de comentarios de revisores de tesis, lo que puede contribuir significativamente al avance de la investigación académica y al mejoramiento de la calidad de las tesis presentadas.

Palabras clave: Comentarios de revisores de tesis, clasificación temática, Inferencia latente, Modelización Latent Dirichlet (LDA), taxonomía.

ABSTRACT

The main objective of this research was to apply the Latent Dirichlet Allocation (LDA) modeling technique as a latent inference approach for thematic classification of comments made by thesis reviewers. We aimed to comprehensively evaluate the capacity of the LDA model to identify and group comments into diverse thematic categories, with the purpose of obtaining a more precise understanding and organization of the information contained in thesis reviews. To achieve this objective, we collected a corpus of data comprising comments from thesis reviewers registered in the Research Platform Integrated to Academic Work with Responsibility (PILAR) of the prestigious Universidad Nacional del Altiplano de Puno (UNAP). Subsequently, we applied the LDA model to generate a classification system that effectively assigned thesis reviews to established thematic categories based on a predefined taxonomy. Finally, we conducted a thorough evaluation of the effectiveness and accuracy of the developed model to determine its optimal capability in classifying comments. The obtained results demonstrate the potential of the LDA technique as an effective tool for thematic classification of thesis reviewers' comments, which can significantly contribute to the advancement of academic research and improvement of the quality of presented theses.

Keywords: Latent Dirichlet Modeling (LDA), latent inference, thematic classification, thesis reviewer comments, taxonomy.

INTRODUCCIÓN

La presente tesis "Inferencia Latente y la Modelización de Dirichlet Latente para la Caracterización de Revisiones de Tesis" tuvo como objetivo principal explorar el uso de la inferencia latente y la modelización de Dirichlet latente como herramientas efectivas para la comprensión y caracterización de las revisiones de tesis. La inferencia latente es un enfoque poderoso en el análisis de datos textuales, ya que permite descubrir patrones ocultos y estructuras subyacentes en grandes conjuntos de texto. En el contexto de las revisiones de tesis, la inferencia latente puede ser especialmente útil para identificar los temas centrales abordados en las tesis y analizar la información contenida en los comentarios de los revisores. Al aplicar la inferencia latente, es posible descubrir categorías temáticas subyacentes y agrupar los comentarios relacionados en clusters significativos. Esto facilita una comprensión más profunda de los aspectos clave de las tesis evaluadas y ayuda a los investigadores a obtener una visión global de los comentarios realizados por los revisores.

La modelización de Dirichlet latente, en particular, es una técnica que ha demostrado ser efectiva en la clasificación temática de documentos textuales. Al utilizar esta técnica, se pueden asignar probabilidades a cada palabra dentro de un documento, lo que permite identificar los temas dominantes en una colección de revisiones de tesis. Esta modelización proporciona una estructura sólida para la organización y análisis de los comentarios, lo que resulta en una comprensión más precisa de las contribuciones y limitaciones de las tesis evaluadas. Durante décadas, las revisiones de tesis han sido una parte fundamental del proceso de evaluación académica. Sin embargo, el manejo y la interpretación de la gran cantidad de información contenida en estas revisiones representa un desafío significativo. Por lo tanto, es fundamental desarrollar métodos efectivos para analizar y clasificar los comentarios de los revisores con el fin de obtener una visión más clara y estructurada de los temas abordados en las tesis.

La técnica de modelización LDA se ha destacado como una herramienta poderosa para la clasificación temática en diferentes dominios. Su capacidad para identificar patrones ocultos en los datos textuales la convierte en una opción prometedora para analizar y caracterizar las revisiones de tesis. La inferencia latente proporcionada por el modelo LDA permite descubrir las categorías temáticas subyacentes presentes en los comentarios de los revisores. Para llevar a cabo este estudio, se recopiló un corpus de datos que consta

de revisiones de tesis registradas en la Plataforma Académica Integrada a la Labor Académica con Responsabilidad (PILAR). Estas revisiones fueron proporcionadas por revisores expertos de las diferentes escuelas profesionales de la UNA PUNO. Posteriormente, se aplicó la técnica de modelización LDA para generar un modelo que asignara automáticamente los comentarios a diferentes categorías temáticas.

Una vez generado el modelo LDA y establecida la taxonomía, se llevó a cabo una evaluación exhaustiva para medir la eficacia y precisión de la clasificación temática. Se utilizaron métricas de evaluación estándar, como la precisión, la exhaustividad y la puntuación F, para determinar la calidad de la clasificación obtenida. A medida que avanzamos en la era digital y nos enfrentamos a grandes cantidades de datos textuales, es esencial contar con herramientas y técnicas efectivas que nos ayuden a extraer conocimientos significativos y comprender mejor la información contenida en las revisiones académicas. En los próximos años, se espera que la investigación se enfoque en la mejora y refinamiento de los modelos existentes de inferencia latente y modelización de Dirichlet latente, con el objetivo de lograr una mayor precisión en la clasificación temática y una comprensión más profunda de las temáticas abordadas en las revisiones de tesis. Además, se espera que se realicen esfuerzos para adaptar y aplicar estas técnicas a diferentes contextos y dominios académicos, permitiendo así su amplia utilización en diversas disciplinas. Asimismo, el desarrollo de nuevas herramientas y algoritmos basados en la inferencia latente y la modelización de Dirichlet latente será fundamental para avanzar en esta área de investigación. Estas herramientas deberán ser capaces de manejar conjuntos de datos cada vez más grandes y complejos, permitiendo un procesamiento eficiente y preciso de las revisiones de tesis.

El presente informe se organiza conforme la especificación de la unidad de postgrado, donde en el Capítulo 1, se presenta el marco teórico y se enuncian algunos antecedentes que se relaciona al tema de estudio; en el Capítulo 2, se sustenta de mejor manera el problema antes expuesto, incluyendo una especificación de la justificación, y la determinación de los objetivos y las hipótesis; en el Capítulo 3: se expone por cada uno de los objetivos como se desarrolló la recolección y tratamiento de datos sobre una muestra de mercados; y en el Capítulo 4, en orden conforme los objetivos se expone todos los resultados que se ha logrado obtener en la investigación, incluyendo una discusión con resultados de otros autores que han desarrollado trabajos casi parecidos; finalmente

de manera breve y concreta se presentan las conclusiones y recomendaciones que se arribó con este estudio.

CAPÍTULO I

REVISIÓN DE LITERATURA

1.1. Marco teórico

1.1.1. Inferencia latente

1.1.1.1. Definición de inferencia latente

La inferencia latente es un concepto fundamental en estadísticas y modelización de datos que juega un papel crucial en la comprensión y análisis de conjuntos de datos complejos. Se refiere al proceso de estimar o inferir características o variables no observables en un conjunto de datos a partir de las variables observables (Savoy *et al.,* 2012). Estas variables no observables, comúnmente denominadas "variables latentes", son aquellas que influyen en los datos observados pero que no son directamente medibles. En esencia, la inferencia latente busca desentrañar lo que yace en la sombra, revelando o descubriendo estas variables ocultas a través de técnicas estadísticas y modelos matemáticos (Priyantina & Sarno, 2019). En el contexto de la investigación científica y el análisis de datos, la inferencia latente permite capturar relaciones subyacentes, patrones ocultos y estructuras no evidentes en los datos observados. Esto puede ser especialmente útil cuando se trata de datos multidimensionales o cuando se buscan simplificar conjuntos de datos complejos para su comprensión o modelado (Martínez-Comeche, 2023; Morchid & Linar`es, 2012). La inferencia latente se ha aplicado en una amplia variedad de campos, desde el procesamiento de lenguaje natural hasta la genética, la psicología y la

economía, desempeñando un papel crucial en la revelación de información valiosa que de otro modo permanecería oculta (Martínez-Comeche, 2023).

Una de las aplicaciones más conocidas de la inferencia latente es la modelización de Dirichlet latente (LDA), que se utiliza para descubrir tópicos ocultos en conjuntos de documentos. Esta técnica permite identificar patrones temáticos en grandes colecciones de texto, lo que facilita la categorización, la extracción de conocimiento y la agrupación de documentos. En resumen, la inferencia latente es una herramienta esencial en la investigación y el análisis de datos que desvela los misterios que residen bajo la superficie de los datos observados, ofreciendo una visión más profunda y completa de los fenómenos estudiados (Gropp *et al.*, 2019).

1.1.1.2. Variables latentes en el análisis de datos

En el análisis de datos, las variables latentes son un concepto fundamental que se utiliza para capturar y modelar características o factores subyacentes que influyen en las variables observadas, pero que no pueden ser directamente medidas o observadas. Estas variables latentes son invisibles en sí mismas, pero su presencia se manifiesta a través de su influencia en las variables que sí podemos medir (Tran *et al.*, 2019). Aquí se presentan cuatro aspectos clave relacionados con las variables latentes en el análisis de datos:

Definición de Variables Latentes

Las variables latentes son constructos teóricos o conceptos abstractos que representan aspectos no observables de un fenómeno. Estas variables no pueden ser medidas directamente con herramientas físicas o instrumentos, pero se asumen o inferirán a partir de observaciones o datos indirectos. Por ejemplo, en psicología, la inteligencia puede considerarse una variable latente que se infiere a partir de pruebas cognitivas observables (Kozlowski *et al.*, 2020).

Aplicaciones en Modelos Estadísticos

Las variables latentes se utilizan comúnmente en modelos estadísticos para capturar la variabilidad no observada en los datos. Los modelos de

regresión, análisis factorial, análisis de componentes principales y modelización de mezcla de distribuciones son ejemplos de técnicas que incorporan variables latentes para explicar la relación entre variables observadas (Kozlowski *et al.*, 2020).

Ejemplos en Diversos Campos

Las variables latentes se encuentran en una variedad de campos. En la economía, por ejemplo, se utilizan en modelos de series temporales para representar factores económicos no observables que influyen en las variables económicas observadas, como el PIB. En la biología, se emplean para modelar genes o proteínas latentes que afectan las expresiones génicas observadas. En el análisis de encuestas, las actitudes latentes pueden ser inferidas a partir de respuestas a preguntas específicas (Kozlowski *et al.*, 2020).

Desafíos y Beneficios

Si bien las variables latentes permiten una modelización más completa y realista de los fenómenos, también plantean desafíos, ya que su naturaleza no observable requiere técnicas de inferencia estadística. Sin embargo, al incorporar variables latentes, los modelos pueden capturar mejor la complejidad de los sistemas subyacentes, lo que a menudo lleva a una comprensión más profunda y precisa de los datos y fenómenos estudiados (Kozlowski *et al.*, 2020).

1.1.1.3. Importancia de la inferencia latente en la investigación

investigaciones, los fenómenos estudiados pueden ser extremadamente complejos y multidimensionales. La inferencia latente permite identificar y modelar las estructuras subyacentes que influyen en los datos observados, lo que proporciona una comprensión más profunda y coherente de los sistemas estudiados. (Rieger *et al.*, 2020).

• **Simplificación de Datos Complejos:** Cuando se trabaja con conjuntos de datos extensos o ricos en detalles, la inferencia latente puede simplificar la información al identificar dimensiones latentes o factores clave. Esto facilita

la visualización, la interpretación y el análisis de los datos, lo que a menudo resulta en conclusiones más claras y procesos de toma de decisiones más eficaces (Rieger *et al.*, 2020).

• **Predicción y Clasificación Mejoradas:** La incorporación de variables latentes en modelos predictivos y de clasificación puede aumentar la precisión y la capacidad de generalización de los modelos. Al considerar factores no observables, se pueden desarrollar modelos más robustos que tengan en cuenta la complejidad subyacente de los fenómenos estudiados. Por ejemplo, en el aprendizaje automático, la inferencia latente es fundamental en técnicas como los modelos de mezcla gaussiana (Rieger *et al.*, 2020).

• **Descubrimiento de Conexiones Causales:** En ciertas investigaciones, la inferencia latente puede ayudar a identificar relaciones causales ocultas entre variables observadas. Al comprender mejor cómo las variables latentes afectan las observaciones, los investigadores pueden abordar preguntas de causa y efecto de manera más precisa, lo que es esencial en campos como la epidemiología y la economía (Rieger *et al.*, 2020).

• **Minería de Texto y Análisis de Sentimiento:** En el procesamiento de lenguaje natural, la inferencia latente es crucial para la detección de tópicos, el análisis de sentimientos y la agrupación de documentos. Permite descubrir temas latentes en grandes conjuntos de texto, lo que es valioso en áreas como la investigación de mercado, la toma de decisiones empresariales y la comprensión de la opinión pública (Rieger *et al.*, 2020).

1.1.2. Evolución de la inferencia latente

La evolución de la inferencia latente ha sido un proceso fascinante que ha transformado radicalmente la forma en que abordamos la comprensión y el análisis de datos complejos a lo largo de la historia. En sus orígenes, la inferencia latente encuentra sus raíces en la estadística bayesiana del siglo XIX, donde figuras notables como Thomas Bayes y Pierre-Simon Laplace sentaron las bases para estimar parámetros desconocidos a partir de datos observados utilizando distribuciones de probabilidad. Esta primera etapa permitió la conceptualización

inicial de variables latentes, donde la incertidumbre inherente en los datos podía ser modelada de manera más precisa (Zhao *et al.*, 2020).

A medida que avanzó el tiempo, surgieron avances significativos en la inferencia latente. El análisis factorial, desarrollado por Charles Spearman, fue uno de los primeros métodos que involucraba la inferencia latente, centrándose en la descomposición de la varianza en factores latentes que influían en múltiples variables observadas. Posteriormente, el análisis de componentes principales (PCA) se convirtió en una técnica ampliamente utilizada para reducir la dimensionalidad de datos, identificando factores latentes en los datos observados. Con el tiempo, se desarrollaron los modelos de mezcla de distribuciones, que permitían representar datos como una combinación de múltiples distribuciones de probabilidad, cada una representando un componente latente o una clase (Kiatkawsin *et al.*, 2020; Zhao et al., 2020). Estos avances sentaron las bases para el análisis y la modelización de datos complejos en una variedad de disciplinas. Sin embargo, la evolución de la inferencia latente no se detuvo ahí. Los modelos ocultos de Markov (HMM) se convirtieron en un hito importante en la inferencia latente, especialmente en el procesamiento del lenguaje natural y el reconocimiento de patrones (Robalino & Stalin, 2019). Estos modelos permitieron capturar secuencias de eventos con estados ocultos, lo que resultó fundamental en aplicaciones como el reconocimiento de voz y la traducción automática (Cheng *et al.*, 2020).

Finalmente, la introducción de los modelos de Dirichlet latente (LDA), desarrollados por David Blei y colaboradores, revolucionó la modelización de tópicos en documentos de texto. Estos modelos emplean la inferencia bayesiana para descubrir tópicos latentes en colecciones de documentos, proporcionando una herramienta esencial en el análisis de datos no estructurados y en la comprensión de la información contenida en grandes conjuntos de texto (Shakeel *et al.*, 2018). En conjunto, la evolución de la inferencia latente representa un viaje intelectual a lo largo de los siglos, desde las primeras incursiones en la estadística bayesiana hasta la sofisticación contemporánea de técnicas como LDA. Esta evolución ha impulsado la investigación en una amplia gama de campos, desde la ciencia de datos hasta la inteligencia artificial y la toma de decisiones, y ha permitido la extracción de información valiosa de datos cada vez más complejos y diversos (Suadaa *et al.*, 2016; Ye *et al.*, 2021).

Estadística Bayesiana y sus Contribuciones

La Estadística Bayesiana es un enfoque fundamental en la teoría estadística que se basa en los principios de la probabilidad bayesiana. A diferencia de la estadística clásica, que utiliza métodos frecuentistas para estimar parámetros y tomar decisiones, la estadística bayesiana se centra en la actualización de creencias a medida que se obtiene nueva información. Esta disciplina tiene sus raíces en los trabajos del reverendo Thomas Bayes y Pierre-Simon Laplace en el siglo XVIII, pero ha experimentado un renacimiento significativo en las últimas décadas debido a los avances tecnológicos y la disponibilidad de métodos computacionales (Liu *et al.*, 2020).

La estadística bayesiana se basa en el teorema de Bayes, que establece cómo actualizar nuestras creencias sobre un evento o un parámetro a medida que se obtiene nueva evidencia. La formulación bayesiana implica el uso de una distribución a priori que representa nuestras creencias iniciales y una función de verosimilitud que describe cómo la evidencia afecta esas creencias (Liu et al., 2020). A medida que se obtiene nueva información, la distribución a priori se actualiza para convertirse en la distribución a posteriori, que es la distribución de probabilidad revisada basada en la información disponible.

Contribuciones clave de la Estadística Bayesiana:

- Flexibilidad en la modelización
- Incorporación de incertidumbre
- Manejo de tamaños de muestra
- Aplicaciones en una variedad de campos
- Desarrollo de métodos computacionales
- Aproximaciones Bayesianas Variacionales
- Predicciones y toma de decisiones

El Papel del Análisis Factorial en la Inferencia Latente

El análisis factorial desempeña un papel esencial en la inferencia latente, una rama fundamental de la estadística y la investigación social. La inferencia latente se refiere a la identificación y medición de constructos que no son directamente observables a partir de variables que sí lo son. Aquí es donde entra en juego el análisis factorial, una técnica estadística que se utiliza para abordar este tipo de problemas (Chen *et al.*, 2019). Su función principal radica en modelar las relaciones subyacentes entre las variables observables y, a partir de esta modelización, extraer información crucial sobre las dimensiones no observables que subyacen a estas variables (Chen *et al.*, 2019; Osmani *et al.*, 2020). Este enfoque es de suma importancia en diversos campos. Por ejemplo, en psicología, el análisis factorial se emplea para identificar dimensiones latentes como la inteligencia, la personalidad o la satisfacción laboral a partir de respuestas a cuestionarios. Además de la identificación de estas dimensiones, el análisis factorial cumple una función crucial al reducir la dimensionalidad de los datos. Esta reducción simplifica datos complejos al agrupar variables correlacionadas en dimensiones latentes, lo que facilita enormemente su interpretación y análisis posteriores (Li *et al.*, 2020).

Una de las aplicaciones más destacadas del análisis factorial es la validación de instrumentos de medición, especialmente en psicometría. Esta técnica permite evaluar si las preguntas agrupadas en cuestionarios realmente miden el constructo subyacente que se pretende evaluar, mejorando así la calidad de las mediciones. Además, el análisis factorial también es utilizado para modelar la estructura de los datos, revelando patrones subyacentes que no son inmediatamente evidentes en los datos originales. Esto es especialmente útil en análisis de mercado, donde puede ayudar a identificar las preferencias del consumidor y las relaciones entre productos a partir de datos de compra (Chauhan & Shah, 2021; Jelodar *et al.*, 2019).

La Influencia del Análisis de Componentes Principales (PCA)

El Análisis de Componentes Principales (PCA) es una técnica estadística ampliamente utilizada que desempeña un papel crucial en el análisis de datos multivariados y en la reducción de la dimensionalidad (Putri *et al.*, 2017). PCA se ha convertido en una herramienta indispensable en campos tan diversos como la estadística, la ingeniería, la biología, la economía y la informática, debido a su capacidad para extraer información importante de conjuntos de datos complejos.

Una de las contribuciones más significativas de PCA es su capacidad para reducir la dimensionalidad de los datos. En situaciones en las que se trabaja con conjuntos de datos que contienen muchas variables interrelacionadas, PCA permite simplificar estos datos al identificar las direcciones (llamadas componentes principales) en las cuales la variabilidad es más significativa. Al eliminar las componentes menos informativas, se obtiene una representación más compacta de los datos sin perder demasiada información. Esto es especialmente valioso en la visualización de datos y la simplificación de modelos estadísticos (Chauhan & Shah, 2021; Putri *et al.,* 2017).

Otra influencia importante de PCA es su capacidad para revelar patrones ocultos en los datos. Al proyectar los datos originales en las direcciones de las componentes principales, se pueden identificar relaciones y estructuras subyacentes que pueden no ser evidentes en una vista inicial de los datos. Esta capacidad de descubrimiento de patrones es fundamental en la investigación científica y en la toma de decisiones informadas en campos como la biología, donde PCA se utiliza para analizar datos de expresión génica y detectar relaciones entre genes (Chauhan & Shah, 2021; Nallapati & Cohen, 2008). Además, PCA también juega un papel crucial en la reducción de la colinealidad en modelos estadísticos. Cuando las variables predictoras están altamente correlacionadas, los modelos pueden volverse inestables y difíciles de interpretar. PCA permite crear nuevas variables no correlacionadas (las componentes principales) que pueden utilizarse como predictores en lugar de las variables originales, lo que mejora la estabilidad y la interpretabilidad de los modelos (Backenroth *et al.*, 2017).

Modelos de Mezcla de Distribuciones como Precursores

Los Modelos de Mezcla de Distribuciones son precursores fundamentales en la estadística y el aprendizaje automático, ya que proporcionan un enfoque poderoso para modelar la heterogeneidad en los datos. Estos modelos se utilizan en una amplia variedad de aplicaciones, desde la clasificación de patrones hasta la segmentación de clientes y la detección de anomalías. Su influencia radica en varias áreas clave (Kang *et al.*, 2019):

- **Modelado de Datos Complejos**

Los datos del mundo real suelen ser complejos y están formados por múltiples subgrupos o poblaciones. Los modelos de mezcla de distribuciones permiten representar estas complejidades al asumir que los datos provienen de una mezcla de varias distribuciones subyacentes. Esto facilita la captura de la heterogeneidad y mejora la precisión de los modelos (Kang *et al.*, 2019).

- **Agrupamiento y Segmentación**

En el análisis de agrupamiento (clustering), los modelos de mezcla de distribuciones se utilizan para asignar observaciones a grupos basados en sus características comunes. Esto es esencial en aplicaciones como la segmentación de clientes en marketing o la detección de clases no etiquetadas en aprendizaje no supervisado (Kang *et al.*, 2019).

- **Detección de Anomalías**

En la detección de anomalías, estos modelos permiten identificar observaciones inusuales que no siguen el patrón de las distribuciones subyacentes. Esto es valioso en la seguridad cibernética, la detección de fraudes y la inspección de calidad, entre otros campos (Kang *et al.*, 2019).

- **Aprendizaje Automático Generativo**

Los modelos de mezcla de distribuciones también forman la base de los modelos generativos en el aprendizaje automático, como el modelo de mezcla gaussiana (GMM) y las redes generativas adversariales (GAN). Estos modelos se utilizan para generar datos sintéticos que se asemejan a los datos de entrenamiento, lo que es útil en la síntesis de imágenes, el procesamiento de lenguaje natural y la generación de datos de entrenamiento aumentados (Kang *et al.*, 2019).

- **Inferencia Bayesiana**

Los modelos de mezcla de distribuciones se pueden abordar desde una perspectiva bayesiana, lo que permite incorporar información previa y actualizar creencias a medida que se obtienen datos adicionales. Esto es especialmente útil en la estadística bayesiana y la toma de decisiones bajo incertidumbre (Kang *et al.*, 2019).

Modelos Ocultos de Markov (HMM) y su Importancia

Los Modelos Ocultos de Markov (HMM, por sus siglas en inglés) son una clase fundamental de modelos estadísticos que se utilizan ampliamente en diversas áreas, desde el procesamiento de señales hasta el procesamiento de lenguaje natural. Su importancia radica en su capacidad para modelar y comprender secuencias de datos, así como en su versatilidad en aplicaciones prácticas. Aquí se describe su importancia (Toubia *et al.*, 2019):

Modelado de Secuencias

Los HMM son especialmente útiles para modelar secuencias de datos, donde la estructura temporal es fundamental. Esto los hace adecuados para aplicaciones que involucran datos secuenciales, como el reconocimiento de voz, el análisis de series temporales financieras y la predicción del tiempo (Arnautov *et al.*, 2016).

1.1.3. Modelos de Dirichlet Latente (LDA)

El Modelo de Dirichlet Latente (LDA, por sus siglas en inglés) es un modelo probabilístico que se utiliza para descubrir patrones de temas latentes en colecciones de documentos textuales. LDA es una técnica de aprendizaje no supervisado que asume que los documentos están compuestos por una mezcla de temas y que cada tema se caracteriza por una distribución de palabras. El objetivo de LDA es descomponer una colección de documentos en estos temas subyacentes, lo que permite analizar y categorizar documentos en función de los temas que contienen. Es ampliamente utilizado en aplicaciones de procesamiento de lenguaje natural, como la categorización de texto, la recomendación de contenido y la extracción de información (Toubia *et al.*, 2019; Webb *et al.*, 2020).

1.1.3.1. Introducción a la Modelización de Dirichlet Latente (LDA)

La Modelización de Dirichlet Latente (LDA) es una técnica poderosa en el procesamiento de lenguaje natural y la minería de textos que se utiliza para descubrir estructuras temáticas en grandes conjuntos de documentos. Este enfoque asume que los documentos se generan a partir de una mezcla de temas subyacentes, donde cada tema se representa como una distribución de palabras. LDA se basa en la idea de que, si comprendemos la distribución

de temas en una colección de documentos, podemos inferir la estructura latente que subyace a los textos. Para lograr esto, LDA utiliza un enfoque probabilístico y bayesiano, lo que significa que modela la incertidumbre inherente en la asignación de palabras a temas y en la estructura misma de los temas. Al aplicar LDA, los analistas pueden extraer conocimientos valiosos, realizar tareas de clasificación de documentos basadas en temas y mejorar la comprensión de grandes conjuntos de datos de texto, lo que lo convierte en una herramienta esencial en la minería de texto y el análisis de datos basados en texto (Wang & Taylor, 2019).

1.1.3.2. Fundamentos Probabilísticos de LDA

Los fundamentos probabilísticos de la Modelización de Dirichlet Latente (LDA) se basan en la idea fundamental de que los documentos en una colección se generan a partir de un proceso probabilístico. LDA utiliza un modelo generativo para representar cómo se crean los documentos y cómo se distribuyen las palabras en esos documentos. Aquí se resumen los fundamentos probabilísticos clave de LDA (Kanungsukkasem & Leelanupab, 2019):

- **Mezcla de Temas**

LDA parte del supuesto de que cada documento es una mezcla de temas subyacentes. Esta mezcla se representa mediante una distribución de probabilidad sobre los temas, lo que significa que un documento puede contener múltiples temas en proporciones diferentes (Qomariyah *et al.*, 2019).

- **Temas como Distribuciones de Palabras**

Cada tema se caracteriza como una distribución de palabras, donde ciertas palabras tienen una probabilidad más alta de aparecer en ese tema. En otras palabras, los temas son representados por distribuciones de probabilidad sobre el vocabulario (Qomariyah *et al.*, 2019).

- **Generación de Documentos**

Para generar un documento en LDA, se elige primero una distribución de temas de acuerdo con las probabilidades especificadas. Luego, para cada palabra en el documento, se selecciona un tema de acuerdo con la distribución de temas y, finalmente, se elige una palabra específica de ese tema. Este proceso se repite para todas las palabras en el documento (Qomariyah *et al.*, 2019).

- **Inferencia Inversa**

El objetivo principal de LDA es realizar una inferencia inversa: a partir de un conjunto de documentos observados, se busca encontrar las distribuciones de temas subyacentes y las distribuciones de palabras que mejor explican esos documentos. Esto se hace utilizando métodos de estimación probabilística, como el muestreo de Gibbs o la variational inference (Qomariyah *et al.*, 2019).

- **Parámetros del Modelo**

LDA tiene varios parámetros importantes, como la distribución a priori de temas en la colección, la distribución a priori de palabras en los temas y las distribuciones de temas en cada documento. La estimación de estos parámetros es esencial para ajustar el modelo a los datos observados (Qomariyah *et al.*, 2019).

Los fundamentos probabilísticos de LDA se centran en la generación probabilística de documentos a partir de una mezcla de temas subyacentes y la posterior inferencia de las distribuciones de temas y palabras que mejor explican los datos observados. Esto permite modelar la estructura latente en colecciones de documentos y es la base de la capacidad de LDA para descubrir temas y realizar análisis de texto en una amplia variedad de aplicaciones (Negara *et al.*, 2019; Toubia *et al.*, 2019).

Aplicaciones en Procesamiento de Lenguaje Natural (NLP) y Minería de Texto

Las aplicaciones de la Modelización de Dirichlet Latente (LDA) en Procesamiento de Lenguaje Natural (NLP) y Minería de Texto son diversas

y abarcan una amplia gama de aplicaciones prácticas. A continuación, se describen algunas de las aplicaciones más destacadas (Syed & Spruit, 2017):

• Agrupación de Documentos: Clasificar documentos en grupos según temas comunes (Syed & Spruit, 2017).

• Etiquetado Automático de Temas: Asignar etiquetas temáticas a documentos de forma automática (Syed & Spruit, 2017).

• Análisis de Sentimientos: Identificar opiniones y temas en textos, como en reseñas de productos (Syed & Spruit, 2017).

• Recomendación de Contenido: Sugerir contenido relevante basado en intereses del usuario (Syed & Spruit, 2017).

• Detección de Anomalías: Identificar documentos o fragmentos inusuales en colecciones (Syed & Spruit, 2017).

• Extracción de Información: Recopilar datos relevantes, como nombres y lugares, de documentos (Syed & Spruit, 2017).

• Clasificación de Texto: Categorizar textos en grupos predefinidos (Syed & Spruit, 2017).

• Generación de Texto: Mejorar la generación automática de texto coherente (Syed & Spruit, 2017).

• Análisis de Contenido Web: Analizar contenido y estructura de sitios web (Syed & Spruit, 2017).

• Traducción Automática y Resumen: Ayudar en la traducción y resumen precisos de texto.

Ventajas de LDA en la Extracción de Tópicos

LDA ofrece varias ventajas significativas en la extracción de tópicos a partir de conjuntos de documentos. En primer lugar, LDA es un método no supervisado, lo que significa que no requiere etiquetas o anotaciones previas de los documentos. Esto lo hace altamente escalable y aplicable a grandes volúmenes de datos sin la necesidad de una intervención manual extensa.

En segundo lugar, LDA permite la identificación de tópicos latentes que pueden no ser evidentes a simple vista. Esto es crucial en el análisis de grandes colecciones de documentos donde la estructura temática subyacente puede ser compleja y diversa (Wahyudi & Kusumaningrum, 2019).

Además, LDA proporciona una representación compacta de los documentos en términos de tópicos, lo que facilita la visualización y el resumen de grandes conjuntos de datos de texto. Esta capacidad de reducir la dimensionalidad de los datos de manera significativa es especialmente valiosa en aplicaciones de procesamiento de lenguaje natural y análisis de texto. Otra ventaja importante del modelo LDA es su capacidad para manejar documentos multitemáticos, es decir, documentos que pueden abordar múltiples temas en lugar de uno solo. Esto refleja la realidad de muchos conjuntos de datos de texto donde la diversidad temática es común. Es importante mencionar que el modelo LDA ofrece ventajas notables en la extracción de tópicos al ser un método no supervisado, capaz de descubrir tópicos latentes, proporcionar representaciones compactas de documentos y manejar documentos multitemáticos, lo que lo convierte en una herramienta valiosa en el análisis de grandes conjuntos de datos de texto (Wahyudi & Kusumaningrum, 2019).

1.1.4. Elementos Clave en la Modelización de Dirichlet Latente (LDA)

La Modelización de Dirichlet Latente (LDA) es un enfoque poderoso en el procesamiento de lenguaje natural y la minería de textos que se utiliza para desentrañar la estructura temática subyacente en grandes conjuntos de documentos. En su esencia, LDA se basa en varios elementos clave que son fundamentales para su funcionamiento (Egger & Yu, 2022). En primer lugar, considera que cada documento es una mezcla de tópicos latentes, lo que significa que los temas subyacentes están presentes en diferentes proporciones en cada documento. Además, LDA representa los tópicos como distribuciones de palabras, lo que implica que cada tópico está compuesto por un conjunto de palabras con probabilidades específicas de ocurrencia. Este enfoque probabilístico permite que LDA modele la generación de documentos a través de un proceso de selección de tópicos y palabras, lo que da como resultado la estructura latente de los documentos.

Para estimar los tópicos y las distribuciones de palabras que mejor explican los datos observados, LDA utiliza técnicas de inferencia estadística, como el muestreo de Gibbs o la inferencia variacional. Estos elementos clave de LDA, como la mezcla de tópicos, las distribuciones de palabras y la inferencia estadística, permiten desvelar las estructuras temáticas ocultas en colecciones de documentos, lo que es fundamental en una amplia gama de aplicaciones, desde la organización de información hasta la recomendación de contenido y el análisis de sentimientos (Egger & Yu, 2022; Wahyudi & Kusumaningrum, 2019).

Documentos como Unidades Fundamentales

En la Modelización de Dirichlet Latente (LDA), los documentos se consideran las unidades fundamentales de análisis. Esta elección refleja la naturaleza del procesamiento de lenguaje natural y la minería de texto, donde los documentos, que pueden ser artículos, libros, reseñas, correos electrónicos u otros tipos de textos, son la unidad básica de información. La razón detrás de considerar los documentos como unidades fundamentales radica en su importancia como contenedores de contenido y la forma en que se estructuran las colecciones de texto. Cada documento se ve como una composición única de palabras y, por lo tanto, se representa como un conjunto de palabras que forman su contenido. La premisa subyacente es que los documentos comparten palabras relacionadas entre sí debido a su contenido común. Por ejemplo, un conjunto de artículos de noticias sobre política tendrá palabras relacionadas como "elección", "gobierno" y "candidato". LDA busca identificar los tópicos subyacentes que explican estas asociaciones de palabras en los documentos (Celard *et al.*, 2020).

Al considerar los documentos como unidades fundamentales, LDA permite el análisis y la agrupación de documentos en función de su contenido temático. Esto tiene aplicaciones significativas en la categorización de documentos, la recuperación de información, la recomendación de contenido y la extracción de información relevante. Además, LDA es capaz de manejar colecciones de documentos de diferentes tamaños y estructuras, lo que lo convierte en una herramienta versátil para el análisis de texto en una amplia variedad de dominios. Al considerar los documentos como las unidades de análisis centrales, LDA permite una comprensión más profunda de la estructura y los patrones temáticos en grandes

conjuntos de datos de texto, lo que es esencial en la era de la información y el análisis de datos basados en texto (Celard *et al.*, 2020; Deveaud *et al.*, 2014).

Tópicos como Conceptos Latentes

En la Modelización de Dirichlet Latente (LDA), los "tópicos" se consideran como conceptos latentes o subyacentes que representan las estructuras temáticas en los documentos. Estos tópicos son fundamentales en la comprensión y el análisis de grandes conjuntos de documentos, ya que capturan las relaciones semánticas y los patrones temáticos que subyacen en el contenido textual. Cada tópico se concibe como una distribución de palabras, lo que significa que está compuesto por un conjunto de palabras con probabilidades específicas de ocurrencia. Por ejemplo, un tópico relacionado con la política podría incluir palabras como "elección", "gobierno", "partido" y "candidato", con probabilidades más altas de aparición para estas palabras dentro de ese tópico en particular (Celard *et al.*, 2020; Nzali *et al.*, 2016).

La idea central es que los documentos se generan a partir de una mezcla de estos tópicos. Cada documento tiene su propia distribución de tópicos, lo que implica que puede tratar diferentes temas en diferentes proporciones. Esta combinación de tópicos en un documento particular permite explicar cómo se compone su contenido temático. Por ejemplo, un artículo de noticias podría ser una mezcla de tópicos relacionados con la política, la economía y el deporte. Los tópicos como conceptos latentes en LDA son esenciales porque permiten modelar la diversidad temática en colecciones de documentos y, al mismo tiempo, proporcionan una representación compacta y estructurada de los temas abordados en los textos. Esto facilita la organización, la clasificación y la comprensión de grandes volúmenes de datos de texto, lo que es esencial en una variedad de aplicaciones, desde la categorización de documentos hasta la recomendación de contenido y el análisis de sentimientos. En resumen, los tópicos son conceptos latentes que representan temas subyacentes en los documentos y desempeñan un papel central en la Modelización de Dirichlet Latente al capturar la esencia temática de los textos (Savoy *et al.*, 2012; Zhou *et al.*, 2021).

Palabras como Componentes de Tópicos

En el contexto de la Modelización de Dirichlet Latente (LDA), las "palabras" son consideradas como los componentes fundamentales de los tópicos. Esta relación entre palabras y tópicos es esencial para comprender cómo LDA modela la estructura temática en colecciones de documentos. En LDA, cada tópico se representa como una distribución de palabras. Esto significa que un tópico en particular está compuesto por un conjunto de palabras que tienen probabilidades específicas de ocurrencia dentro de ese tópico. Estas probabilidades indican cuán relevantes son las palabras para ese tópico en particular. Por ejemplo, en un tópico relacionado con "ciencia", las palabras como "experimento", "hipótesis" y "descubrimiento" tendrían probabilidades más altas de aparición, mientras que palabras como "fútbol" o "receta" tendrían probabilidades más bajas. La relación entre palabras y tópicos permite a LDA modelar cómo se generan los documentos. Cuando se crea un documento, se selecciona una mezcla de tópicos y, a partir de esta mezcla, se eligen las palabras específicas en el documento. Esto refleja la idea de que los documentos son una combinación de tópicos, y las palabras dentro de esos documentos son indicativas de los tópicos presentes (Alrumiah & Al-Shargabi, 2022; Zhou *et al.*, 2021).

El enfoque de LDA en las palabras como componentes de tópicos es esencial para descubrir y entender las estructuras temáticas en los documentos. Permite que LDA asigne palabras relevantes a tópicos específicos y, al mismo tiempo, captura cómo los tópicos se mezclan en los documentos reales (Zhang *et al.*, 2017). Esto proporciona una representación efectiva de los documentos en términos de temas, lo que facilita la organización y el análisis de grandes conjuntos de datos de texto en aplicaciones como la categorización de documentos, la extracción de información y la generación de resúmenes, entre otras. En resumen, las palabras son los componentes clave que permiten a LDA modelar los tópicos y, por lo tanto, la estructura temática en los documentos de texto (Alrumiah & Al-Shargabi, 2022; Zhang *et al.*, 2017).

1.1.5. Proceso de Inferencia con Modelización de Dirichlet Latente (LDA)

El proceso de inferencia en la Modelización de Dirichlet Latente (LDA) es una etapa fundamental que permite estimar los tópicos y las distribuciones de palabras que mejor explican un conjunto de documentos observados. En este proceso, se

utiliza información estadística y probabilística para descubrir la estructura latente subyacente en los datos de texto (Bastani *et al.*, 2018). La inferencia en LDA se basa en la idea de que, dado un conjunto de documentos, se pueden inferir los tópicos y las asignaciones de palabras a tópicos más probables que generaron esos documentos. A través de técnicas como el muestreo de Gibbs o la inferencia variacional, LDA estima estos parámetros de manera eficiente, lo que permite un análisis profundo de la composición temática de los documentos y la extracción de información valiosa a partir de grandes conjuntos de datos textuales. En este contexto, exploraremos el proceso de inferencia en LDA y su importancia en la revelación de patrones temáticos ocultos en los documentos (Calvo-Valverde *et al.*, 2020; Martínez-Comeche, 2023).

Inicialización de Parámetros en LDA

La inicialización de parámetros en la Modelización de Dirichlet Latente (LDA) es un paso crucial antes de llevar a cabo el proceso de inferencia para estimar los tópicos y las distribuciones de palabras en un conjunto de documentos (Cazalens *et al.*, 2013). Aunque LDA es un modelo probabilístico, necesita valores iniciales para sus parámetros antes de converger hacia las estimaciones finales. Aquí se explica brevemente cómo funciona la inicialización de parámetros en LDA:

- **Número de Tópicos**

Uno de los parámetros clave en LDA es el número de tópicos que se desea identificar en los documentos. Esta cantidad debe establecerse antes de iniciar la inferencia y puede ser determinada de antemano o mediante técnicas de validación cruzada (Daud *et al.*, 2018).

- **Distribución de Tópicos en Documentos**

Antes de la inferencia, es común asignar una distribución inicial de tópicos a cada documento en la colección. Esto se puede hacer de manera aleatoria o utilizando información previa, si está disponible. Por ejemplo, se podría suponer que todos los documentos contienen una mezcla uniforme de tópicos en una inicialización aleatoria (Daud *et al.*, 2018).

- **Distribución de Palabras en Tópicos**

Similarmente, se asigna una distribución inicial de palabras a cada tópico. Estas distribuciones pueden ser generadas de manera aleatoria o basadas en conocimientos previos. Por ejemplo, se podría asignar una distribución inicial equitativa de palabras a cada tópico (Daud *et al.*, 2018).

- **Estimación Iterativa**

Con estas inicializaciones en su lugar, el proceso de inferencia iterativa en LDA comienza para ajustar las distribuciones de tópicos y palabras de manera que se ajusten mejor a los datos observados. Esto implica actualizaciones repetidas de las asignaciones de tópicos a palabras en cada documento y las estimaciones de las distribuciones de tópicos y palabras (Daud *et al.*, 2018).

La inicialización adecuada de parámetros es esencial en LDA, ya que puede afectar significativamente la convergencia del modelo y la calidad de las estimaciones finales de tópicos y palabras. Una inicialización pobre puede llevar a resultados subóptimos, mientras que una inicialización adecuada puede acelerar la convergencia y mejorar la calidad de los resultados. Por lo tanto, la elección de los valores iniciales de los parámetros en LDA es una consideración importante en su implementación y aplicación efectiva (Daud et al., 2018).

Elección del Número de Tópicos (K)

La elección del número de tópicos (K) en la Modelización de Dirichlet Latente (LDA) es una decisión crítica que tiene un impacto significativo en la interpretación y el rendimiento del modelo. Determinar el valor óptimo de K es un desafío en sí mismo y no existe una regla fija para seleccionarlo. En su lugar, se utiliza un enfoque basado en evaluación y comprensión del dominio de aplicación (Hoblos, 2020). Aquí se describen algunas consideraciones clave al elegir K:

La elección del número de tópicos (K) en la Modelización de Dirichlet Latente (LDA) es un aspecto crucial en la implementación efectiva de este modelo de procesamiento de lenguaje natural. K determina cuántos tópicos distintos se intentarán descubrir en la colección de documentos y, por lo tanto, afecta directamente la calidad de las inferencias realizadas. Sin embargo, no existe una regla fija para seleccionar K, ya que su elección depende en gran medida del contexto y los objetivos específicos del análisis. En este sentido, exploraremos más

a fondo las consideraciones clave y las estrategias que se utilizan comúnmente para abordar esta importante decisión al trabajar con LDA (Hoblos, 2020; Negara *et al.*, 2019).

- **Exploración de Datos**

Explorar los datos y comprender el contexto de su colección de documentos. Ayuda significativamente a identificar la diversidad temática y dar una idea inicial del número de tópicos potenciales (Hoblos, 2020; Negara *et al.*, 2019).

- **Métodos de Evaluación**

Utilice métodos de evaluación como la perplejidad perplexity, la coherencia de tópicos y la validación cruzada para comparar el rendimiento del modelo LDA en diferentes valores de K. La perplejidad tiende a disminuir a medida que K aumenta, pero la coherencia de tópicos puede ayudar a identificar un punto óptimo (Hoblos, 2020; Negara *et al.*, 2019).

- **Balance**

Busque un equilibrio entre tener suficientes tópicos para capturar la diversidad temática en los documentos y no tener demasiados tópicos que dificulten la interpretación. Un número excesivo de tópicos puede generar fragmentación y redundancia en los resultados (Hoblos, 2020; Negara *et al.*, 2019).

- **Aplicación Específica**

Considere la aplicación específica para la que está utilizando LDA. Algunas aplicaciones pueden requerir un mayor grado de detalle temático (un valor mayor de K), mientras que otras pueden funcionar bien con una representación más general (un valor menor de K) (Hoblos, 2020; Negara *et al.*, 2019).

- **Validación Cruzada**

Utilice técnicas de validación cruzada para evaluar el rendimiento del modelo LDA en datos de prueba independientes. Esto puede ayudar a evitar el sobreajuste y garantizar que el modelo generalice bien (Hoblos, 2020; Negara *et al.*, 2019).

- **Experimentación Iterativa**

Se realizó diferentes experimentos iterativos con diferentes valores de K y evaluar cómo afectan la calidad de los tópicos y la coherencia del modelo (Hoblos, 2020; Negara *et al.*, 2019).

En última instancia, la elección de K en LDA es un proceso iterativo que combina la intuición del dominio, la evaluación cuantitativa y la comprensión cualitativa de los resultados. No hay un valor único "correcto" de K, ya que depende de las características específicas de los datos y los objetivos de la aplicación. El objetivo es encontrar un valor de K que equilibre la capacidad del modelo para capturar la estructura temática en los documentos con la interpretabilidad y la utilidad práctica de los tópicos identificados (Hoblos, 2020; Negara *et al.*, 2019).

Asignación Aleatoria de Tópicos a Palabras

La asignación aleatoria de tópicos a palabras es un paso fundamental en el proceso de inicialización de la Modelización de Dirichlet Latente (LDA) y es esencial para comenzar la inferencia y estimación de tópicos y distribuciones de palabras. Este proceso implica asignar tópicos a cada palabra en cada documento de manera aleatoria antes de que el modelo comience a ajustar las asignaciones de tópicos de manera iterativa (Negara *et al.*, 2019; Wang *et al.*, 2019). En esta fase de inicialización, cada palabra en un documento se asigna aleatoriamente a uno de los K tópicos posibles, donde K es el número de tópicos predeterminado por el usuario. Esta asignación aleatoria puede llevar a resultados iniciales bastante incoherentes, ya que no se basa en ninguna información sobre las relaciones reales entre palabras y tópicos en los documentos. Sin embargo, esta inicialización es esencial para dar inicio al proceso de inferencia y, posteriormente, al ajuste de las asignaciones de tópicos de acuerdo con los patrones de co-ocurrencia observados en los datos reales (Wang *et al.*, 2019).

Una vez que se ha realizado la asignación aleatoria inicial, el modelo LDA comienza a ajustar estas asignaciones iterativamente en función de la estructura temática que se encuentra en los documentos observados. A través de múltiples pasos de muestreo y optimización, el modelo LDA busca mejorar la coherencia de las asignaciones de tópicos, lo que resulta en una representación más precisa de la estructura temática de la colección de documentos. A pesar de que la asignación aleatoria inicial puede parecer rudimentaria, es un paso crucial en el proceso de

inferencia de LDA. Permite que el modelo comience desde una base neutral y se adapte gradualmente a los patrones de tópicos reales en los datos. A medida que avanza la inferencia, las asignaciones de tópicos se ajustan para reflejar mejor la estructura subyacente de los documentos, lo que finalmente conduce a una representación más precisa de los tópicos y palabras en la colección de texto (Negara *et al.*, 2019; Toubia *et al.*, 2019; R. Wang *et al.*, 2019).

Creación de una Asignación Inicial de Tópicos

La creación de una asignación inicial de tópicos en la Modelización de Dirichlet Latente (LDA) es un paso crítico para iniciar el proceso de inferencia y estimación de tópicos y distribuciones de palabras en un conjunto de documentos. Esta asignación inicial proporciona un punto de partida desde el cual el modelo LDA ajustará gradualmente las asignaciones de tópicos a palabras en función de la estructura temática real presente en los documentos (Wang *et al.*, 2020). Aquí se describen las etapas clave en la creación de una asignación inicial de tópicos:

- **Número de Tópicos (K)**

Antes de crear la asignación inicial, debe decidirse el número de tópicos (K) que se desea identificar en la colección de documentos. Este valor puede determinarse según la exploración de datos, la comprensión del dominio y las necesidades específicas de la aplicación (Deveaud *et al.*, 2014; Wang *et al.*, 2020).

- **Asignación Aleatoria**

Para cada palabra en cada documento, se asigna aleatoriamente uno de los K tópicos posibles como su tópico inicial. Esta asignación aleatoria proporciona un punto de partida neutral, aunque no informativo, y es esencial para evitar sesgos iniciales (Deveaud *et al.*, 2014; Y. Wang *et al.*, 2020).

- **Iteraciones de Inferencia**

A partir de esta asignación inicial aleatoria, el modelo LDA inicia el proceso de inferencia. A través de iteraciones repetidas, el modelo ajusta las asignaciones de tópicos en función de la probabilidad condicional y la estructura temática observada en los datos reales (Deveaud *et al.*, 2014; Y. Wang *et al.*, 2020).

- **Optimización de Parámetros**

Durante la inferencia, LDA también optimiza otros parámetros, como las distribuciones de tópicos en documentos y las distribuciones de palabras en tópicos, para maximizar la probabilidad conjunta de los datos observados (Deveaud *et al.*, 2014; Wang *et al.*, 2020).

La creación de una asignación inicial de tópicos aleatoria es esencial para evitar que el modelo LDA comience con suposiciones previas que puedan sesgar los resultados. A medida que avanza la inferencia, las asignaciones de tópicos se ajustan para reflejar mejor los patrones reales de co-ocurrencia de palabras y tópicos en la colección de documentos. Esto resulta en una representación más precisa de la estructura temática subyacente en los textos y es un paso crucial en el proceso de modelización de LDA (Deveaud *et al.*, 2014; Wang *et al.*, 2020).

Estimación de Distribución de Tópicos en Documentos

La estimación de la distribución de tópicos en documentos es una etapa esencial en la Modelización de Dirichlet Latente (LDA) y tiene como objetivo determinar cómo se distribuyen los tópicos en cada uno de los documentos de la colección (Billami *et al.*, 2020; Xue *et al.*, 2020). Este proceso es fundamental para entender la composición temática de los textos y revelar las relaciones entre los documentos y los tópicos subyacentes. La estimación comienza con una asignación inicial aleatoria de tópicos a palabras en cada documento. Luego, mediante métodos de inferencia como el muestreo de Gibbs o la inferencia variacional, el modelo LDA ajusta gradualmente estas asignaciones para reflejar la estructura temática real presente en los datos observados. A medida que progresa la inferencia, se calcula la distribución de tópicos en cada documento, que muestra la probabilidad de que cada tópico esté presente en ese documento específico. Esta distribución se representa como un vector en el que cada componente corresponde a un tópico y su valor indica la probabilidad de que el tópico esté presente en el documento (Grisales-Aguirre & Figueroa-Vallejo, 2022).

La estimación de la distribución de tópicos en documentos tiene aplicaciones prácticas en diversas áreas, como la categorización de documentos, la recuperación de información y la recomendación de contenido. Permite organizar y entender la

diversidad temática en una colección de documentos, lo que facilita la búsqueda y el análisis de información relevante. Además, es esencial para la identificación de documentos relevantes en sistemas de recomendación basados en tópicos y para realizar análisis de tendencias y cambios temáticos en conjuntos de datos textuales a lo largo del tiempo. En resumen, la estimación de la distribución de tópicos en documentos es un paso crucial en LDA que revela cómo se relacionan los tópicos con los documentos y proporciona información valiosa en diversas aplicaciones de procesamiento de lenguaje natural y minería de texto (Grisales-Aguirre & Figueroa-Vallejo, 2022).

Estimación de Distribución de Palabras en Tópicos

La estimación de la distribución de palabras en tópicos es otro aspecto fundamental en la Modelización de Dirichlet Latente (LDA) y tiene como objetivo determinar cómo se distribuyen las palabras dentro de cada uno de los tópicos identificados en la colección de documentos. Este proceso es crucial para comprender la esencia y el contenido de cada tópico y para obtener una representación significativa de la estructura temática subyacente en los textos (Grisales-Aguirre & Figueroa-Vallejo, 2022).

La estimación comienza con una asignación inicial aleatoria de tópicos a palabras en todos los documentos. A medida que avanza el proceso de inferencia, el modelo LDA ajusta las asignaciones de tópicos de manera iterativa para reflejar la estructura temática real en los datos observados. Al mismo tiempo, se calcula la distribución de palabras en cada tópico, que muestra las probabilidades de que cada palabra en el vocabulario esté asociada con ese tópico específico. Esta distribución de palabras en tópicos se representa como un vector, donde cada componente corresponde a una palabra en el vocabulario y su valor indica la probabilidad de que esa palabra esté relacionada con el tópico en cuestión. Como resultado, se obtiene un conjunto de distribuciones de palabras para cada uno de los tópicos identificados en la colección de documentos (Grisales-Aguirre & Figueroa-Vallejo, 2022).

La estimación de la distribución de palabras en tópicos es esencial para interpretar y etiquetar los tópicos de manera significativa. Permite identificar las palabras clave que caracterizan cada tópico y proporciona información sobre su contenido temático. Además, esta información es fundamental para aplicaciones como la

generación de resúmenes de texto, la clasificación de documentos en categorías temáticas y la búsqueda de información relevante. En resumen, la estimación de la distribución de palabras en tópicos es un paso crítico en LDA que permite capturar la esencia de los tópicos y su relación con las palabras en la colección de documentos, lo que facilita una comprensión más profunda y útil de la estructura temática de los textos (Grisales-Aguirre & Figueroa-Vallejo, 2022).

Identificación de Palabras Representativas de Tópicos

La identificación de palabras representativas de tópicos es una tarea clave en la Modelización de Dirichlet Latente (LDA) y se refiere a la selección de las palabras más relevantes y distintivas que caracterizan un tópico específico (Blei *et al.*, 2003; Hoblos, 2020). Este proceso tiene como objetivo proporcionar una descripción coherente y significativa de cada tópico identificado en la colección de documentos. Aquí se explican los pasos involucrados en la identificación de palabras representativas de tópicos:

- **Distribución de Palabras en Tópicos**

Antes de identificar las palabras representativas, es necesario estimar la distribución de palabras en cada tópico. Esto se logra durante el proceso de inferencia de LDA, donde se calculan las probabilidades de que cada palabra en el vocabulario esté asociada con un tópico en particular (Hoblos, 2020).

- **Selección de las Palabras Principales**

Una vez que se tiene la distribución de palabras en tópicos, se pueden seleccionar las palabras más relevantes y distintivas para cada tópico. Esto se hace comúnmente tomando las palabras con las probabilidades más altas en la distribución de palabras de ese tópico en particular (Hoblos, 2020).

- **Filtrado de Palabras Comunes**

Para mejorar la calidad de las palabras representativas, es común filtrar palabras comunes, como artículos y preposiciones, que no aportan información distintiva sobre el tópico (Hoblos, 2020).

- **Interpretación y Etiquetado**

Después de seleccionar las palabras representativas, es importante interpretarlas y, si es necesario, etiquetar el tópico en función de estas palabras clave. Esto ayuda a comprender mejor la esencia temática del tópico y a facilitar su identificación (Hoblos, 2020).

- **Visualización y Análisis**

Finalmente, las palabras representativas se utilizan para visualizar y analizar los tópicos en un formato comprensible. Esto puede incluir la creación de gráficos, nubes de palabras o resúmenes temáticos que muestren las palabras clave y la importancia relativa de cada tópico (Hoblos, 2020).

La identificación de palabras representativas de tópicos es esencial para la interpretación y la utilidad de los resultados de LDA. Proporciona una descripción clara y coherente de los tópicos identificados, lo que facilita la comprensión de la estructura temática subyacente en una colección de documentos. Además, estas palabras representativas son útiles en una variedad de aplicaciones, como la categorización de documentos, la búsqueda de información y la recomendación de contenido, donde se utilizan para etiquetar y organizar eficazmente el contenido textual (Hoblos, 2020).

Proceso Iterativo de Ajuste de Distribuciones

El proceso iterativo de ajuste de distribuciones en la Modelización de Dirichlet Latente (LDA) es una fase crítica que tiene como objetivo estimar de manera precisa las distribuciones de tópicos en documentos y las distribuciones de palabras en tópicos en una colección de documentos (Hoblos, 2020). Este proceso se lleva a cabo en varias etapas bien definidas para lograr convergencia y obtener resultados significativos:

- **Inicialización**

El proceso comienza con una asignación inicial aleatoria de tópicos a las palabras en cada documento. Esta asignación aleatoria sirve como punto de partida neutral, sin suposiciones previas sobre la estructura temática de los documentos (Hoblos, 2020).

- **Iteraciones**

A continuación, se inicia un ciclo de iteraciones. En cada iteración, el modelo LDA revisa y ajusta las asignaciones de tópicos en función de las palabras y los patrones de co-ocurrencia observados en los documentos. La cantidad de iteraciones puede variar, pero es común realizar varias pasadas hasta que se alcance la convergencia (Hoblos, 2020).

- **Actualización de Distribuciones**

Durante cada iteración, se actualizan las distribuciones de tópicos en documentos y las distribuciones de palabras en tópicos. Esto se hace utilizando información sobre las asignaciones de tópicos actuales en los documentos y las probabilidades de co-ocurrencia de palabras con tópicos específicos (Hoblos, 2020).

- **Convergencia**

El proceso iterativo continúa hasta que se alcanza la convergencia, lo que significa que las distribuciones de tópicos y palabras han convergido hacia estimaciones estables y representativas de la estructura temática en los documentos. La convergencia se verifica mediante criterios predefinidos, como la estabilidad de las estimaciones a lo largo de varias iteraciones (Hoblos, 2020).

- **Resultados Finales**

Al final del proceso iterativo, se obtienen las estimaciones finales de las distribuciones de tópicos en documentos y las distribuciones de palabras en tópicos. Estas distribuciones representan la estructura temática identificada en la colección de documentos y se utilizan para analizar, categorizar y comprender los textos en función de sus temas subyacentes (Hoblos, 2020).

El proceso iterativo de ajuste de distribuciones en LDA es esencial para garantizar que las estimaciones sean precisas y coherentes con los datos reales. A medida que avanza la inferencia, el modelo se adapta y mejora continuamente las asignaciones de tópicos, lo que resulta en una representación más fiel de la estructura temática en los textos. Este enfoque iterativo es una de las fortalezas de LDA y lo hace especialmente útil en aplicaciones de procesamiento de lenguaje natural y minería de texto (Hoblos, 2020).

Convergencia del Modelo LDA

La convergencia del modelo LDA (Modelización de Dirichlet Latente) es un aspecto crítico en el proceso iterativo de estimación de tópicos y distribuciones de palabras (Celard *et al.*, 2020; Y. Wang et al., 2020). La convergencia se refiere al estado en el que el modelo ha alcanzado estimaciones estables y representativas de las distribuciones de tópicos y palabras en la colección de documentos. Para monitorear la convergencia, es importante utilizar diversos criterios, como la estabilidad de las estimaciones a lo largo de las iteraciones, la disminución de la perplejidad (una métrica comúnmente utilizada), y la convergencia de las distribuciones de tópicos y palabras (Chen *et al.*, 2019).

Uno de los indicadores clave de convergencia es la estabilidad de las estimaciones de tópicos y palabras a lo largo de las iteraciones. A medida que el modelo LDA converge, las distribuciones de tópicos y palabras tienden a dejar de cambiar significativamente en cada iteración, lo que indica que se han estabilizado. La perplejidad es otra métrica relevante que disminuye constantemente a medida que avanzan las iteraciones, lo que sugiere que el modelo está mejorando su capacidad para predecir las palabras en los documentos, un indicador de convergencia. Además de estos criterios, también es importante observar la convergencia de las distribuciones de tópicos en documentos y las distribuciones de palabras en tópicos. Cuando estas distribuciones convergen hacia valores estables, las estimaciones finales son coherentes y representativas de la estructura temática subyacente en los documentos. Las evaluaciones empíricas, como la coherencia de tópicos y la interpretación de las palabras representativas de tópicos, también pueden ayudar a verificar si el modelo ha convergido satisfactoriamente (Wang *et al.*, 2020).

Criterios para Determinar la Convergencia

Los criterios para determinar la convergencia en la Modelización de Dirichlet Latente (LDA) son fundamentales para evaluar si el proceso iterativo ha alcanzado estimaciones estables y representativas de las distribuciones de tópicos y palabras en la colección de documentos (Wang *et al.*, 2020). A continuación, se enumeran y desarrollan los principales criterios utilizados para determinar la convergencia en LDA:

- **Estabilidad de las Estimaciones**

La estabilidad de las estimaciones es un indicador clave de convergencia. Se refiere a la consistencia de las distribuciones de tópicos y palabras a lo largo de las iteraciones. Cuando las estimaciones de tópicos y palabras dejan de cambiar significativamente en cada iteración, es un signo de que el modelo está convergiendo hacia resultados estables (Wang *et al.*, 2020).

- **Disminución de la Perplejidad (Perplexity)**

La perplejidad es una métrica comúnmente utilizada para evaluar la calidad del modelo LDA y verificar su convergencia. Representa la medida de cuán bien el modelo puede predecir las palabras en los documentos. Una disminución constante en la perplejidad a medida que avanzan las iteraciones sugiere que el modelo está mejorando su capacidad para ajustarse a los datos observados, lo que es un indicador de convergencia (Wang *et al.*, 2020).

- **Convergencia de Distribuciones**

Otro criterio importante es la convergencia de las distribuciones de tópicos en documentos y las distribuciones de palabras en tópicos. Cuando estas distribuciones convergen hacia valores estables, significa que las estimaciones son coherentes y representativas de la estructura temática subyacente en los documentos (Wang *et al.*, 2020).

- **Evaluación Empírica**

Además de los criterios anteriores, las evaluaciones empíricas juegan un papel crucial en la determinación de la convergencia. Esto implica analizar la coherencia de los tópicos, la interpretación de las palabras representativas de tópicos y la utilidad de los resultados en aplicaciones específicas. Si los tópicos son interpretables y útiles para los objetivos de análisis, es un indicio de que el modelo ha convergido satisfactoriamente (Wang *et al.*, 2020).

- **Criterios de Parada**

A menudo, se establecen criterios de parada predefinidos, como un número máximo de iteraciones o una mejora mínima en las métricas de convergencia, para determinar cuándo detener el proceso iterativo. Estos criterios son útiles para evitar

el sobreajuste y garantizar que el modelo no continúe ajustándose innecesariamente (Wang *et al.*, 2020).

Los criterios para determinar la convergencia en LDA son esenciales para garantizar que el modelo haya alcanzado resultados confiables y coherentes. La combinación de criterios como la estabilidad de las estimaciones, la disminución de la perplejidad, la convergencia de distribuciones y la evaluación empírica proporciona una sólida base para determinar cuándo el proceso iterativo ha llegado a un estado convergente y las estimaciones son representativas de la estructura temática en la colección de documentos (Wang *et al.*, 2020).

1.2. Antecedentes

El presente trabajo de investigación no cuenta con una gran variedad de antecedentes debido a que es un tipo de trabajo poco desarrollado, por tal razón se consideran los siguientes antecedentes.

En la investigación desarrollada por Ayuso Luengo en 2022, se propone un enfoque basado en el modelado de tópicos para extraer los temas principales presentes en distintas publicaciones sin la necesidad de leerlas todas. Para lograrlo, se utilizaron tres algoritmos con el fin de encontrar los temas más coherentes. El algoritmo más utilizado en la modelización de tópicos en un conjunto de textos es LDA, el cual ha demostrado ser muy efectivo, pero también presenta ciertas limitaciones. Para cubrir la limitación que LDA tiene al trabajar con documentos de longitud corta, se propone aplicar también BTM, un algoritmo diseñado específicamente para este tipo de textos. (Ayuso, 2022).

En el estudio de Zhou Ya *et al.* (2020), se destaca la importancia de las reseñas para los consumidores en línea al momento de realizar compras. Sin embargo, también se reconoce que existen reseñas spam que no brindan información precisa y confiable sobre los productos. Para abordar esta problemática, se plantea un método basado en el modelo de tópicos y el grado de anomalía del revisor. Este método divide las reseñas spam en dos tipos: aquellas con contenido falso y aquellas que son engañosas. En primer lugar, se modela el conjunto de datos experimental utilizando el modelo de tópicos LDA para detectar las reseñas spam de contenido falso con diferentes temas. Luego, las reseñas spam engañosas se detectan mediante el índice de grado de anomalía del revisor. Este índice asigna una puntuación a cada reseña según las características extraídas y los pesos

relacionados. Finalmente, se combina esta puntuación con un cálculo de peso adaptativo basado en el período anormal y la similitud del revisor para obtener la puntuación de la reseña. Las reseñas con una puntuación alta se consideran spam, mientras que las reseñas con una puntuación baja son consideradas verdaderas. Los experimentos muestran que este método tiene una mejora significativa en la tasa de reconocimiento de reseñas spam (Ya *et al.*, 2020).

En el estudio de Shao, Ding y Li (2021), se destaca el surgimiento de diversos modelos de clasificación de sentimientos basados en redes neuronales profundas en los últimos años. Sin embargo, muchos de estos modelos existentes se entrenan en función de la incrustación de palabras, o dependen de una costosa anotación a nivel de palabras, o solo utilizan anotaciones a nivel de oraciones. Se reconoce que algunos fenómenos lingüísticos y recursos importantes aún no han sido completamente estudiados. Con el objetivo de abordar el fenómeno lingüístico de que una oración puede tener múltiples sentimientos y diferentes palabras objetivo pueden tener diferentes sentimientos, se propone un modelo de clasificación de sentimientos multifunción basado en LDA. El modelo extrae automáticamente las palabras objetivo mediante LDA, filtra las características de sentimiento global de las oraciones, extrae las características de sentimiento local de las oraciones con el vocabulario de sentimiento externo e integra diversas características con el modelo de clasificación de sentimientos. Una serie de experimentos en tres conjuntos de datos demuestran que el modelo multifunción es efectivo. La introducción de LDA no solo puede reducir la demanda de palabras objetivo etiquetadas, mejorar la precisión de la clasificación de sentimientos, sino también analizar con mayor precisión la tendencia emocional interna de los eventos de opinión pública (Shao *et al.*, 2021).

Por otro lado, también se pudo evidenciar que el Modelado de Tópicos es un método popular para extraer el conocimiento semántico oculto presente en una colección de documentos. Debido al aumento en los datos textuales, los modelos de tópicos juegan un papel significativo para inferir temas significativos en textos. En este trabajo, presentamos una revisión de varias representaciones de temas y modelos de tópicos utilizados para descubrir temas en textos largos y cortos. Los métodos de modelado de temas se agrupan en métodos de modelado de temas estándar, métodos basados en agrupamiento, métodos autoagregantes y métodos basados en aprendizaje profundo. Los modelos de temas convencionales como el Modelo de Espacio Vectorial (VSM), el Análisis Semántico Latente (LSA), el Análisis Semántico Latente Probabilístico (PLSA),

la Asignación de Dirichlet Latente (LDA) y la Mezcla Multinomial (MM) se discuten con sus méritos y limitaciones. También se discuten los modelos de tópicos basados en agrupamiento y los modelos de tópicos autoagregantes utilizados para el corpus de texto corto. Se discuten los métodos basados en aprendizaje profundo con el modelado de temas tradicionales para mejorar la extracción de temas de alta calidad (Yamunathangam et al., 2021).

Hussain *et al.*, (2023) utilizó un enfoque de minería de texto para descubrir las dimensiones de calidad del servicio de banca móvil (m-banking) a partir de las revisiones de los clientes sobre las aplicaciones de m-banking. Se aplicó el método de asignación latente de Dirichlet (LDA) para descubrir estas dimensiones en las revisiones positivas y negativas de 24 bancos que operan en Pakistán. Se encontró que las revisiones positivas se centraban en la seguridad, la conveniencia, la facilidad de uso, la mejora continua, la utilidad y los atributos de la aplicación, mientras que las revisiones negativas se enfocaban en la disponibilidad del sistema, la capacidad de respuesta, las actualizaciones defectuosas, los problemas de inicio de sesión y la fiabilidad. Este estudio proporciona información útil para mejorar la experiencia del servicio de m-banking para los clientes y también demuestra cómo los gerentes pueden emplear técnicas analíticas de texto para evaluar y mejorar la calidad de los servicios (Hussain *et al.*, 2023).

Christy & Meera Gandhi conceptualizan que en la era digital, la modelización de temas (topic modeling) es fundamental para la agrupación de documentos. Cada documento se representa como una colección de palabras y asignar distribuciones de palabras a temas específicos es un desafío. Los algoritmos no supervisados utilizados en la modelización de temas no siempre se corresponden con la realidad del mundo. Para solucionar este problema, existen dos métodos: utilizar un algoritmo supervisado que asigne etiquetas a los temas o utilizar un conjunto predefinido de temas y ajustarlos a la distribución de palabras. Ambos métodos presentan inconvenientes, como la asignación de una etiqueta que no tenga significado semántico o la falta de conocimiento para detectar temas en un conjunto cerrado de temas. Este artículo presenta un nuevo algoritmo llamado Concept-Latent Dirichlet Allocation (Concept-LDA) que incorpora el aprendizaje por refuerzo para mejorar la precisión de los temas y de la etiquetación de temas. Los experimentos muestran que Concept-LDA logra una mayor precisión que LDA en la modelización (Christy & Meera, 2019).

Por otro lado, en el artículo de Basilio et al. se desarrolló un método para descubrir conocimiento en bases de datos de respuesta a emergencias basado en informes de incidentes policiales, generando información que identifique las demandas criminales locales para seleccionar el portafolio adecuado de estrategias policiales para solucionar el problema. Se utiliza una metodología de descubrimiento de conocimiento que involucra técnicas de minería de texto utilizando Latent Dirichlet Allocation (LDA) integrado con el método multicriterio ELECTRE I. El método desarrollado permitió la identificación de las demandas criminales más comunes que ocurrieron en las áreas policiales estudiadas en 2016. Se construyeron ocho escenarios diferentes donde se puede identificar que para cada conjunto específico de demandas criminales hay un conjunto de estrategias policiales a aplicar. La metodología desarrollada contribuye a la identificación de prácticas criminales y sus características basadas en informes de ocurrencias policiales almacenados en bases de datos de respuesta a emergencias. La originalidad del estudio radica en la integración de técnicas de minería de texto, LDA y el método ELECTRE I para detectar el delito en una ubicación específica basada en informes de delitos almacenados en bases de datos de respuesta a emergencias, permitiendo la identificación y elección de estrategias policiales personalizadas para demandas criminales específicas (Basilio *et al.,* 2020).

Masood *et al.*, (2017) tuvo como propósito proponer un modelo novedoso para la recomendación de etiquetas en redes sociales llamado MFS-LDA que resuelve el problema de inicio en frío (cold start problem). La mayoría de los métodos de recomendación de etiquetas existentes se basan en etiquetas populares, pero fallan al enfrentar nuevos recursos que aún no están etiquetados. MFS-LDA es un modelo basado en la Latent Dirichlet Allocation (LDA) que utiliza múltiples espacios de características (título, contenido y etiquetas) para recomendar etiquetas. El uso de múltiples espacios de características permite a MFS-LDA recomendar etiquetas incluso si faltan datos de un espacio de características. Los resultados muestran una mejora significativa en comparación con los métodos existentes (Masood *et al.*, 2017).

En aprendizaje automático, los modelos de temas se utilizan para descubrir estructuras abstractas en grandes colecciones de documentos. El artículo desarrollado por (Härdle & Chen, 2017), presenta una selección adaptada de conceptos tanto de la teoría de la información como de la estadística para construir una base sólida para comprender los modelos de temas. Se enfoca en dos modelos en particular: Latent Dirichlet Allocation y

Dynamic Topic Model. Las aplicaciones construidas en el lenguaje de programación Python demuestran posibles casos de aplicación, como encontrar artículos de noticias similares en contenido y explorar la evolución de temas de artículos de noticias con el tiempo. El objetivo de este artículo es guiar al lector desde una comprensión casual de conceptos estadísticos básicos, como los adquiridos típicamente en estudios de pregrado, hacia una comprensión de los modelos de temas (Härdle & Chen, 2017).

El articulo desarrollado por Drissi *et al.*, (2022) presenta una propuesta de clasificación de texto automatizada basada en el algoritmo de modelado de temas LDA y la estructura semántica de los documentos. Se destaca que las consultas de los usuarios son cada vez más complejas y necesitan buscar por tema o documento, no solo por palabras clave. Se realizan experimentos para comparar la eficacia de esta solución con enfoques de clasificación basados solo en el contenido del texto y se demuestra la superioridad de la propuesta presentada (Drissi *et al.*, 2022).

El artículo presentado por (Ya *et al.*, 2020) presenta un método basado en un modelo de temas y el grado de anormalidad del revisor para detectar revisiones de spam en línea, que son importantes para los consumidores a la hora de realizar compras en línea. El método divide las revisiones de spam en dos tipos: basadas en el contenido y engañosas. Primero, se utiliza un modelo de temas LDA para detectar las revisiones de spam basadas en el contenido. Luego, se utilizan medidas de la anormalidad del revisor para detectar las revisiones engañosas. Las revisiones con puntuaciones más altas se consideran spam, mientras que las de puntuaciones más bajas se consideran auténticas. Los experimentos muestran una mejora en la tasa de reconocimiento de las revisiones de spam (Ya *et al.*, 2020).

CAPÍTULO II

PLANTEAMIENTO DEL PROBLEMA

2.1. Identificación del problema

La clasificación temática de los comentarios de revisores de tesis universitarias es una tarea importante en la evaluación y mejora de la calidad de la investigación. Sin embargo, el proceso manual de clasificación es tedioso y susceptible a sesgos que se pueden cometer de parte de los evaluadores o revisores, lo que requiere necesariamente un enfoque automatizado (Nery *et al.*, 2020). La inferencia latente mediante la modelización Latente Dirichlet (LDA) es una técnica ampliamente usada en la de la minería de texto que nos permite identificar temas latentes en un corpus de texto (Drissi *et al.*, 2022). En este sentido, es posible utilizar la modelización LDA para clasificar temáticamente los comentarios de revisores de tesis universitarias.

En el contexto de la gestión de investigación en la Universidad Nacional del Altiplano de Puno, uno de los principales problemas que se enfrenta es la falta de eficiencia en el procesamiento de grandes volúmenes de datos, en particular en el procesamiento de los comentarios de revisores de tesis universitarias. La cantidad de comentarios de revisores de tesis es enorme y la clasificación manual es una tarea difícil y tediosa que consume mucho tiempo y recursos. Este problema puede afectar la calidad del proceso de evaluación de tesis, ya que puede retrasar la toma de decisiones y dificultar la identificación de patrones y tendencias en los comentarios de los revisores. Por lo tanto, es necesario buscar soluciones que permitan una mejor y más eficiente clasificación y caracterización de los comentarios de los revisores, y una posible solución es la utilización de técnicas de inferencia latente y modelización de Dirichlet latente (LDA) para la caracterización automática de las revisiones de tesis.

Otro problema que se enfrenta en la caracterización automática de las revisiones de tesis es la dificultad para capturar la esencia de los comentarios de los revisores. La clasificación temática automática a menudo falla en capturar la esencia de los comentarios debido a la naturaleza subjetiva del lenguaje utilizado en las revisiones de tesis. La identificación de los temas y conceptos subyacentes de revisiones es fundamental para una evaluación adecuada (Poushneh & Rajabi, 2022), ya que permite detectar tendencias y patrones en los comentarios de los revisores y así mejorar el proceso de evaluación (Park & Kwon, 2013). Por lo tanto, es necesario encontrar soluciones que permitan una caracterización automática más precisa y efectiva de los comentarios de los revisores, y una posible solución es la utilización de técnicas de inferencia latente y modelización de Dirichlet latente (LDA) para la identificación de los temas y conceptos subyacentes en las revisiones de tesis.

También es importante denotar las limitaciones a las que nos enfrentamos en la caracterización automática de las revisiones de tesis sobre todo en la precisión de la clasificación temática automática. Es frecuente en la clasificación temática automática puede verse influenciada por factores como el uso de palabras clave específicas o la falta de contexto (Torres-Cruz *et al.*, 2022), lo que puede llevar a una clasificación inexacta o errónea de los comentarios de los revisores. Esto puede afectar la calidad de la evaluación de las tesis y la toma de decisiones por parte de los responsables de la gestión de investigación. Por lo tanto, es necesario buscar soluciones que permitan una clasificación temática más precisa y efectiva de los comentarios de los revisores, y una posible solución es la utilización de técnicas de inferencia latente y modelización de Dirichlet latente (LDA) para la identificación de los temas y conceptos subyacentes en las revisiones de tesis, lo que permitiría una clasificación más precisa y contextualizada.

2.2. Enunciados del problema

El alcance de esta investigación se centra en la aplicación de la técnica de modelización Latente Dirichlet (LDA) para la clasificación temática de los comentarios de revisores de tesis universitarias. Se limita a la recopilación y análisis de revisiones de tesis almacenadas en la Plataforma de Investigación Integrada a la Labor Académica con Responsabilidad (PILAR) de la Universidad Nacional del Altiplano de Puno (UNAP). El estudio se enfoca en evaluar la capacidad de LDA para identificar y agrupar comentarios en diferentes categorías temáticas con el propósito de mejorar la comprensión y

organización de la información en estas revisiones. El alcance no abarca la revisión de tesis de otras instituciones ni la consideración de factores externos que puedan influir en el proceso de revisión.

2.3. Justificación

La evaluación de tesis es un proceso crucial en la gestión de investigación en las instituciones académicas, ya que garantiza la calidad y relevancia de la investigación realizada. Por otro lado, varios autores coinciden en que la evaluación de tesis es fundamental para asegurar la pertinencia, calidad y rigor científico de las investigaciones, por lo que es importante prestar una especial atención a este proceso y por ende al contenido de las revisiones de evaluación (Arias & Giraldo, 2011; de-Migel, 2010). Una evaluación adecuada no solo garantiza la calidad de la investigación, sino que también contribuye al desarrollo de la comunidad científica y al fortalecimiento de la reputación de la institución donde se albergan los trabajos

La revisión de tesis es un proceso crucial para garantizar la calidad y coherencia de los trabajos finales de investigación. La clasificación temática de los comentarios de los revisores permite una evaluación más eficiente y precisa de los trabajos finales.

Sin embargo, la clasificación temática manual es un proceso tardado y sujeto a sesgos humanos, lo que puede resultar en una evaluación no objetiva y poco precisa de los trabajos finales.

El proceso de evaluación manual de tesis presenta varias limitaciones, tales como la falta de eficiencia en el procesamiento de grandes volúmenes de datos, la dificultad para capturar la esencia de los comentarios y la limitación en la precisión de la clasificación temática. Según Yang et al. (2016), "el proceso de evaluación manual de tesis es costoso en términos de tiempo y recursos, y puede ser influenciado por factores subjetivos y personales de los evaluadores, lo que puede afectar la calidad y objetividad de la evaluación". Además, la naturaleza subjetiva del lenguaje hace que la clasificación temática sea una tarea difícil y a menudo poco precisa

Una solución para abordar estas limitaciones es la utilización de técnicas de inferencia latente y modelización de Dirichlet latente (LDA) para la caracterización automática de las revisiones de tesis. Según (Drissi *et al.,* 2022) "LDA es una herramienta efectiva para descubrir patrones subyacentes y temáticas en grandes colecciones de documentos,

utilizando la recuperación de la información permite una clasificación temática automática y precisa". La utilización de estas técnicas permitiría una evaluación más eficiente y efectiva, reduciendo el costo en términos de tiempo y recursos.

El presente trabajo contribuirá, por un lado, la mejora del proceso de evaluación de tesis contribuiría al fortalecimiento de la reputación de la institución y al desarrollo de la comunidad científica. Además, la utilización de técnicas de minería de textos y de inteligencia artificial pueden tener aplicaciones en otros campos, como la búsqueda de información, la identificación de patrones y tendencias, y la detección de fraudes, entre otros. Como mencionan (Hassani et al., 2020) "la aplicación de técnicas de minería de textos en el campo de la investigación académica es una tendencia en crecimiento, y el proyecto de tesis puede contribuir a la generación de nuevos conocimientos y aplicaciones en este campo.

2.4. Objetivos

2.4.1. Objetivo general

Evaluar la eficacia de la modelización Latent Dirichlet (LDA) en la identificación de temas relevantes y recurrentes en las revisiones de tesis universitarias.

2.4.2. Objetivos específicos

- Implementar la modelización Latent Dirichlet (LDA) en las revisiones de tesis universitarias.
- Analizar la capacidad de la modelización LDA para identificar temas recurrentes y relevantes en las revisiones de tesis universitarias.
- Evaluar la precisión y coherencia de los temas identificados por la modelización LDA.

2.5. Hipótesis

2.5.1. Hipótesis general

La utilización de técnicas de inferencia latente y modelización de Dirichlet latente (LDA) permitirá caracterizar de manera eficiente, precisa y efectiva las revisiones

de tesis, proporcionando una mayor comprensión y análisis de la información presente en estas revisiones.

CAPÍTULO III

MATERIALES Y MÉTODOS

3.1. Lugar de estudio

El lugar de estudio es la Universidad Nacional del Altiplano de Puno (UNAP), una institución académica que a la fecha cuenta con sistemas de gestión de la información en investigación por lo que la UNAP proporciona un entorno adecuado para llevar a cabo esta investigación, ya que alberga una cantidad significativa de revisiones de tesis que servirán como fuente de datos.

3.2. Población

La población objetivo de esta investigación se centra en las revisiones de tesis universitarias realizadas en la Universidad Nacional del Altiplano de Puno. Estas revisiones representan un conjunto significativo de comentarios de revisores relacionados con la tesis de pregrado.

3.3. Muestra

La muestra se compondrá de un subconjunto representativo de las revisiones de tesis universitarias de la UNAP. Para garantizar la representatividad de la muestra, se utilizará una técnica de muestreo estratificado que tomará en cuenta diversas áreas temáticas y tendencias de investigación presentes en la universidad. Esta muestra se seleccionará de manera cuidadosa y siguiendo criterios específicos para asegurar la calidad y pertinencia de los datos recopilados.

3.4. Método de investigación

Tipo de Investigación

El presente estudio se enmarca en una investigación de tipo exploratorio y descriptivo. En primera instancia, se explorará la aplicación de la modelización Latent Dirichlet (LDA) en la clasificación temática de revisiones de tesis universitarias, considerando su viabilidad y potencial. Posteriormente, se llevará a cabo una descripción detallada de los resultados obtenidos a partir de la aplicación de esta técnica.

Diseño de Investigación

El diseño de investigación adoptado es no experimental y transversal. Se recopilarán datos de revisiones de tesis almacenadas en la Plataforma de Investigación Integrada a la Labor Académica con Responsabilidad (PILAR) de la Universidad Nacional del Altiplano de Puno (UNAP) en un punto específico en el tiempo. Además, se utilizará un diseño de muestreo estratificado para seleccionar una muestra representativa de revisiones de tesis que abarque diferentes áreas temáticas y tendencias de investigación.

Técnica

La técnica principal utilizada en esta investigación es la modelización Latent Dirichlet (LDA). LDA es un método de procesamiento de lenguaje natural y minería de texto que se emplea para descubrir temas latentes en un corpus de texto. Se aplicará LDA para identificar y agrupar automáticamente los comentarios de revisores de tesis en categorías temáticas, lo que permitirá una caracterización más efectiva de las revisiones.

Instrumento

El principal instrumento utilizado en esta investigación será un sistema de procesamiento de lenguaje natural, como Python, con bibliotecas específicas para la implementación de LDA. Además, se utilizará un equipo de computación adecuado con suficiente capacidad de procesamiento y memoria para llevar a cabo los cálculos necesarios.

Procesamiento de Datos

El procesamiento de datos involucra la recopilación de revisiones de tesis de la Plataforma de Investigación Integrada a la Labor Académica con Responsabilidad (PILAR) de la Universidad Nacional del Altiplano de Puno (UNAP). Estos datos se someterán a un proceso de limpieza y preprocesamiento para eliminar información redundante y ruido. Posteriormente, se aplicará el modelo LDA para la identificación de temas en los comentarios de revisores.

Análisis de Datos

El análisis de datos se centrará en la identificación de los temas latentes en las revisiones de tesis mediante LDA. Además, se llevará a cabo un análisis de la eficacia de esta técnica en la identificación de temas relevantes y recurrentes. Se utilizará la prueba de Análisis de Varianza (ANOVA) para evaluar la hipótesis y comparar los resultados obtenidos con otros métodos o grupos de comparación.

Frecuencia de Términos

La frecuencia de términos se refiere a la cantidad de veces que ciertas palabras o términos específicos aparecen en las revisiones de tesis. Esta información se utilizará en el análisis de datos para identificar los términos más relevantes y representativos de cada tema.

Matriz de documentos-términos

La matriz de documentos-términos es una representación tabular de los documentos (en este caso, las revisiones de tesis) y los términos que aparecen en ellos. Esta matriz se construirá como parte del proceso de preprocesamiento de datos y se utilizará como entrada para la aplicación de la técnica LDA.

Fuente Primaria de Datos

La fuente primaria de datos para esta investigación son las revisiones de tesis almacenadas en la Plataforma de Investigación Integrada a la Labor Académica con Responsabilidad (PILAR) de la Universidad Nacional del Altiplano de Puno (UNAP). Estos datos representan los comentarios de revisores de tesis en un contexto académico y científico.

Tratamiento de Datos

El tratamiento de datos incluye el proceso de limpieza y preprocesamiento de las revisiones de tesis, la construcción de la matriz de documentos-términos y la aplicación de la técnica LDA. Además, se llevará a cabo un análisis estadístico para evaluar la eficacia de la técnica en la identificación de temas.

3.5. Descripción detallada de métodos por objetivos específicos

Operacionalización de Variables

Tabla 1

Operacionalización de variables

Variable	Definición	Operacionalización	Medida
Tema relevante	Temas identificados por la modelización LDA en las revisiones de tesis universitarias.	Aplicación de LDA para identificar temas en comentarios de revisores. Comparación de resultados de LDA con diferentes números de tópicos.	Porcentaje de acierto en la identificación de temas relevantes comparado con otros métodos de identificación de temas.
Tema recurrente	Temas que aparecen con frecuencia en las revisiones de tesis universitarias.	Aplicación de LDA para identificar temas en comentarios de revisores.	Porcentaje de acierto en la identificación de temas recurrentes comparado con otros métodos de identificación de temas.
Eficacia de la modelización LDA	Capacidad de la modelización LDA para identificar temas relevantes y recurrentes en las revisiones de tesis universitarias.	Aplicación de LDA en las revisiones de tesis para identificar temas. Evaluación de la precisión de los temas mediante LDA.	

Recursos Informáticos

En el marco de esta investigación, se contó con una infraestructura informática robusta y actualizada que facilitó la implementación de modelos de procesamiento de lenguaje natural. Se utilizaron computadoras equipadas con hardware de alto rendimiento, lo que

permitió llevar a cabo tareas de procesamiento intensivas en términos de recursos, como la aplicación de la modelización Latent Dirichlet (LDA). Además, se hizo uso de software especializado en procesamiento de lenguaje natural, incluyendo bibliotecas de Python, como Gensim y NLTK, que fueron esenciales para la implementación y ejecución de los modelos. La elección de estas herramientas se basó en su amplia adopción y capacidad para el análisis de texto.

Implementación de Modelos

La implementación de modelos, en particular la aplicación de la modelización Latent Dirichlet (LDA), se realizó siguiendo una metodología estructurada. Se preprocesaron las revisiones de tesis universitarias, incluyendo la tokenización, eliminación de stopwords y otras técnicas de limpieza de texto. Posteriormente, se procedió a la implementación de LDA utilizando la biblioteca Gensim de Python. Se ajustaron los parámetros del modelo, como el número de tópicos, para optimizar su rendimiento. La modelización LDA se aplicó a las revisiones de tesis seleccionadas de acuerdo con el diseño de muestreo estratificado.

Evaluación de Modelos

La evaluación de modelos se llevó a cabo mediante la comparación de los resultados obtenidos con la modelización LDA con respecto a otros métodos de identificación de temas. Se utilizó la prueba de Análisis de Varianza (ANOVA) para determinar si existían diferencias significativas en la capacidad de LDA para identificar temas relevantes y recurrentes en comparación con otros métodos.

Además, se calcularon medidas de evaluación específicas, como el porcentaje de acierto en la identificación de temas relevantes y recurrentes, para medir la eficacia de la modelización LDA. Estas medidas proporcionaron una evaluación cuantitativa de la calidad de los resultados y la capacidad de LDA para cumplir con los objetivos de la investigación. Los resultados de esta evaluación respaldaron la hipótesis planteada en la investigación.

CAPÍTULO IV

RESULTADOS Y DISCUSIÓN

En esta investigación, se evaluó la eficacia de la modelización Latent Dirichlet (LDA) en la identificación de temas relevantes y recurrentes en las revisiones de tesis universitarias. Utilizamos la perplejidad como métrica principal para medir la calidad de los modelos LDA. Se empleó el análisis de regresión por mínimos cuadrados para investigar la relación entre el número de tópicos incorporados en la modelización y dos variables de interés: la perplejidad del modelo y el tiempo de computación requerido. Se estimó un modelo lineal para cada variable, permitiendo evaluar la efectividad de la inclusión de diferentes cantidades de tópicos en la caracterización de las revisiones de tesis. Los resultados revelan una tendencia interesante: a medida que incrementamos el número de tópicos en el modelo LDA, la perplejidad disminuye significativamente. Por ejemplo, con 5 tópicos, la perplejidad fue de aproximadamente -8.40, mientras que, con 20 tópicos, la perplejidad alcanzó un valor de alrededor de -12.35. Esto sugiere una mejora sustancial en el ajuste del modelo a los datos a medida que se aumenta la complejidad de este.

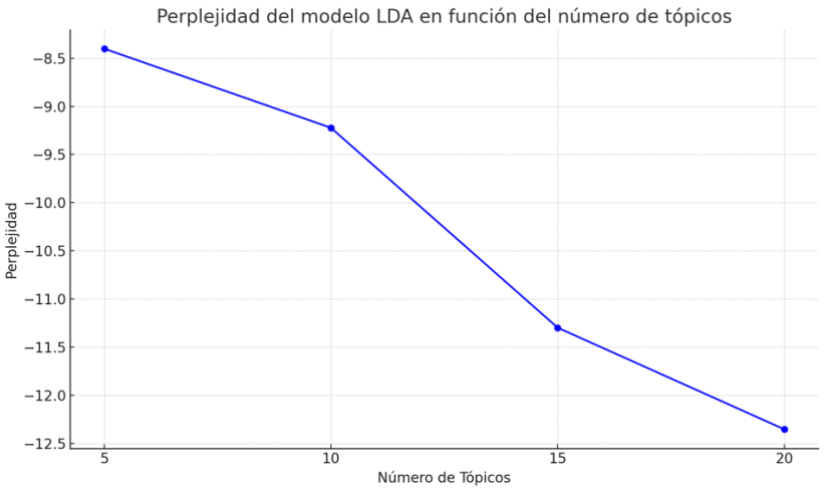

Figura 1. Perplejidad del modelo LDA según número de tópicos.

Por otro lado los resultados indican una correlación negativa entre el número de tópicos y la perplejidad del modelo. Con un coeficiente de -0.2786 y un valor p significativo (0.014), el modelo demuestra que al incrementar el número de tópicos, la perplejidad disminuye, lo cual sugiere una mejor especificación del modelo. La robustez de este resultado se refleja en un valor de R^2 ajustado de 0.959, lo que implica que el modelo explica un 95.9% de la variabilidad en la perplejidad, destacando una fuerte relación entre las variables involucradas.

En contraste, el tiempo de computación mostró una tendencia opuesta. A medida que aumenta el número de tópicos, el tiempo necesario para la caracterización se incrementa, como lo demuestra el coeficiente positivo de 0.0956 y un valor p menor a 0.001. Este hallazgo, con un R^2 ajustado de 0.997, indica que casi toda la variabilidad en el tiempo de computación puede ser explicada por el número de tópicos elegidos para el modelo. Esto destaca una compensación inherente entre la precisión del modelo y su eficiencia computacional.

Entonces es importante destacar que la perplejidad por sí sola no debe ser el único criterio para la selección del número de tópicos. Además de la perplejidad, se analizó el tiempo de ejecución de los modelos LDA en función del número de tópicos. Los resultados también reflejaron una tendencia clara: el tiempo de ejecución aumenta con el incremento en el número de tópicos. Por ejemplo, con 5 tópicos, el tiempo de ejecución fue de aproximadamente 228 segundos, mientras que con 20 tópicos, el tiempo de ejecución se extendió a alrededor de 315 segundos. Este aumento en el tiempo de entrenamiento es un

factor relevante, especialmente en aplicaciones prácticas donde se manejan grandes conjuntos de datos o se necesita una alta eficiencia en el proceso de modelado.

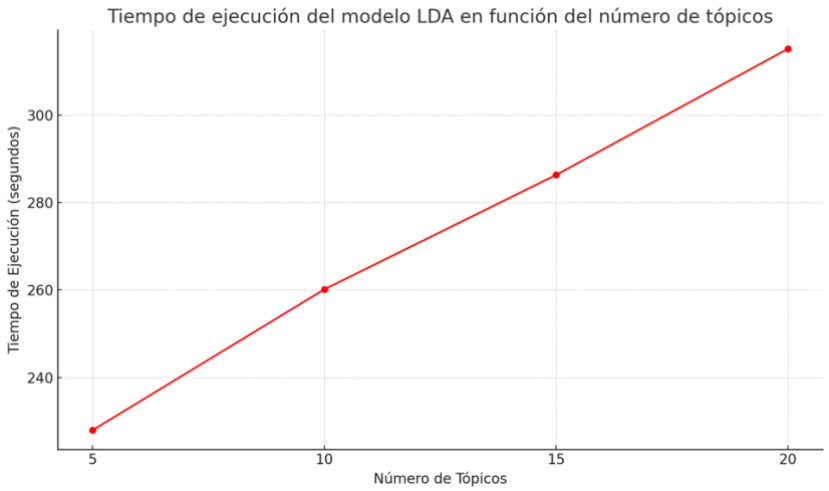

Figura 2. Evaluación del Tiempo en función del número de Tópicos

Tabla 2

Eficacia y Eficiencia del Modelo LDA

N°	Número de Tópicos	Perplejidad	Tiempo (minutos)
1	5	-8.401678	3.7991658051808675
2	10	-9.222946	4.335803882281985
3	15	-11.298803	4.772014498710632
4	20	-12.354300	5.252342716852824

La investigación desarrollada ha revelado diferentes tópicos distintos que corresponden a diferentes etapas o aspectos del proceso de investigación. Estos van desde la conceptualización y justificación del problema hasta la redacción y reflexión sobre el trabajo realizado. Estos resultados subrayan la importancia de abordar cada etapa de la investigación con claridad y precisión, ya que cada una tiene su propia dinámica y enfoque. La capacidad del modelo LDA para identificar estos tópicos subyacentes ofrece una herramienta valiosa para analizar y mejorar el proceso de investigación en su conjunto.

Figura 3. Términos Relevantes en 5 Tópicos

Los resultados del modelo LDA se han clasificado en cinco categorías temáticas que reflejan etapas y aspectos clave del proceso de investigación. El "Tópico 1" aborda la etapa inicial de propuestas y conceptos de proyectos de investigación, mientras que el "Tópico 2" se centra en la revisión de literatura y la construcción de bases teóricas. El "Tópico 3", aunque más ambiguo, parece relacionarse con las técnicas y herramientas metodológicas empleadas en la investigación. El "Tópico 4" es crucial ya que se centra en la identificación y justificación del problema de investigación. Por último, el "Tópico 5" refleja el proceso de escritura, mejora y reflexión del documento final.

Del mismo modo el análisis de los 10 tópicos sugiere la multidimensionalidad del proceso de investigación. Por ejemplo, el "Tópico 1" parece hacer referencia a la estructura y contexto general de un trabajo de investigación, mientras que el "Tópico 2" aborda el proceso de investigación en sí. El "Tópico 3" se enfoca en aspectos más técnicos, posiblemente relacionados con anexos y materiales suplementarios. El "Tópico 4" y el "Tópico 5" abordan aspectos relacionados con la edición, el formato, la redacción y la corrección del documento. Por otro lado, el "Tópico 6" está relacionado con propuestas y observaciones específicas a un contexto local, como una ciudad o región. Estas etiquetas ofrecen una panorámica completa de las múltiples facetas que involucra la realización de una investigación académica o científica.

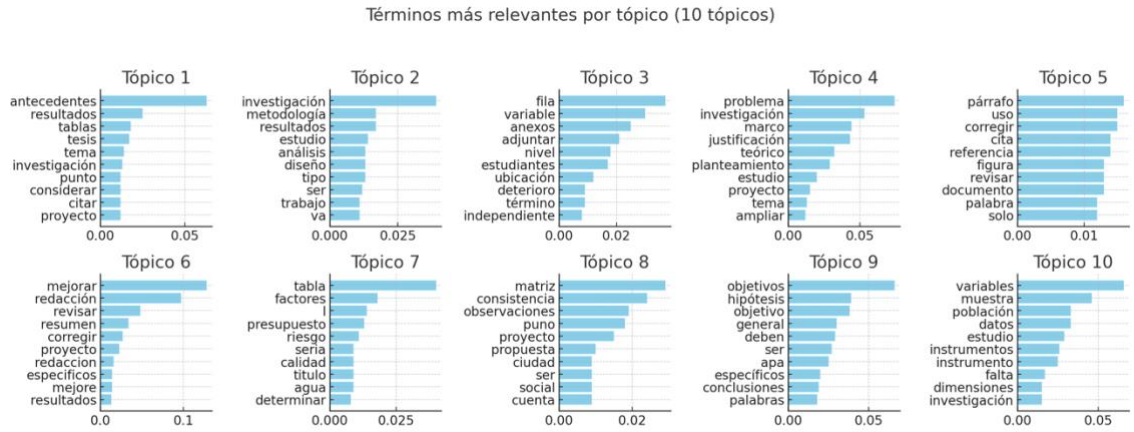

Figura 4. Terminos Relevantes en 10 Tópicos.

Los tópicos proporcionados de la figura sigueinte también abarcan una amplia gama de aspectos relacionados con la investigación académica y científica. Desde elementos textuales básicos, como palabras y términos clave, hasta aspectos más complejos, como metodologías y diseño de investigaciones. También se reflejan preocupaciones sobre el formato, la citación y la presentación del trabajo. Algunos tópicos se centran en el contexto y la ubicación del estudio, mientras que otros abordan la estructura y la organización del proyecto. Las etiquetas proporcionadas ofrecen una visión clara y estructurada de los múltiples componentes y fases que componen un proyecto de investigación, facilitando su análisis y comprensión.

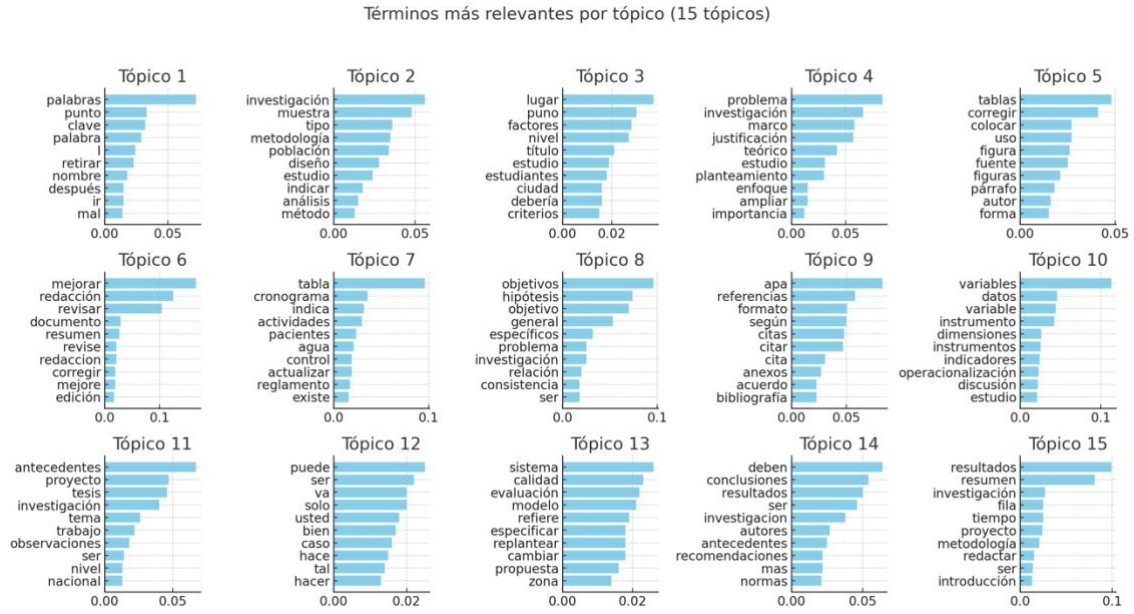

Figura 5. Términos Relevantes en 15 Tópicos.

Los tópicos y términos presentados reflejan aspectos esenciales de la investigación académica y científica. Se abordan desde detalles técnicos, como la metodología y el diseño del estudio, hasta aspectos más generales como la redacción, la ortografía y la estructuración de la investigación. Estos tópicos también muestran la importancia de la claridad y la precisión en la presentación de los resultados y discusiones, así como la necesidad de adherirse a normas específicas (por ejemplo, APA) para citaciones y referencias. En conjunto, estos tópicos subrayan la amplitud y profundidad de consideraciones necesarias para llevar a cabo y presentar un proyecto de investigación de manera coherente y efectiva.

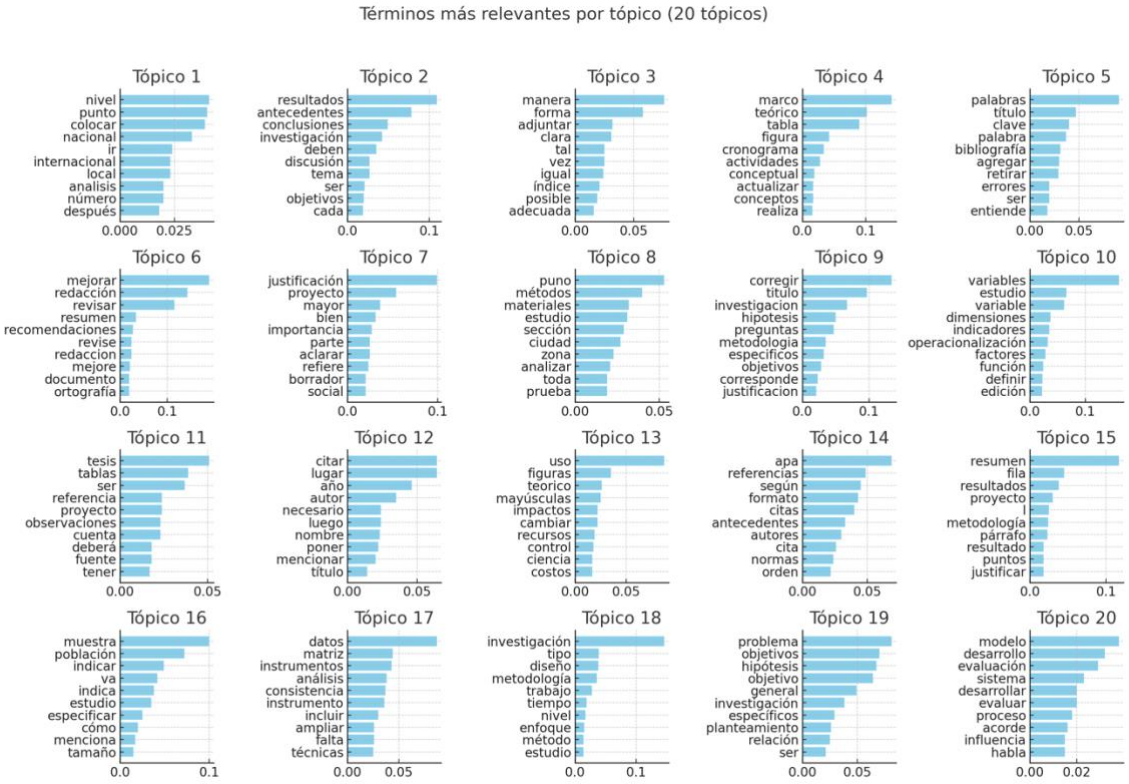

Figura 6. Términos Relevantes en 20 Tópicos.

Tabla 3

Matriz de Conceptualización de tópicos

N°	5 Tópicos	10 Tópicos	15 Tópicos	20 Tópicos
1	Propuesta de Investigación	Estructura y Contexto	Elementos Textuales	Niveles y Ubicación
2	Fundamentación Teórica	Proceso de Investigación	Diseño y Metodología	Resultados y Discusión

N°	5 Tópicos	10 Tópicos	15 Tópicos	20 Tópicos
3	Metodología y Técnica	Aspectos Técnicos y Anexos	Contexto y Ubicación	Claridad y Definición
4	Definición del Problema	Fundamentación del Problema	Marco y Justificación	Marco, Tablas y Cronograma
5	Redacción y Mejora	Edición y Formato	Formato y Presentación	Título y Bibliografía
6		Redacción y Corrección	Edición y Revisión	Redacción y Ortografía
7		Aspectos Organizativos	Organización y Planificación	Justificación y Relevancia
8		Aspectos Locales y Propuestas	Objetivos y Consistencia	Métodos y Zona de Estudio
9		Definición y Objetivos	Normas y Citación	Correcciones y Estructura
10		Metodología y Datos	Variables e Instrumentos	Variables y Medición
11			Contexto del Proyecto	Referencias y Consistencia
12			Comentarios y Observaciones	Uso y Formateo
13			Modelos y Evaluación	Normas APA y Citación
14			Síntesis y Recomendaciones	Resumen y Contenido
15			Resultados y Resumen	Muestra y Población
16				Instrumentos y Análisis
17				Diseño de Investigación
18				Problema y Objetivos
19				Modelo y Evaluación
20				Desarrollo y Aplicación

En el marco de la investigación realizada, se abordaron diferentes aspectos que contribuyeron a la comprensión y contextualización del proyecto. "Propuesta de Investigación", "Estructura y Contexto", "Elementos Textuales" y "Niveles y Ubicación" jugaron un papel fundamental en la conceptualización y el contexto del estudio, proporcionando una base sólida para el desarrollo de la investigación.

La fundamentación teórica y el problema de investigación se exploraron a través de las secciones tituladas "Fundamentación Teórica", "Proceso de Investigación", "Diseño y Metodología", y "Resultados y Discusión". Estas secciones se centraron en la base teórica que sustentó el estudio y en el diseño metodológico empleado para abordar el problema

de investigación. Además, "Definición del Problema" y "Fundamentación del Problema" resaltaron la importancia de identificar y definir claramente el problema que se abordaría en la investigación.

En cuanto a los detalles técnicos y metodológicos, se analizaron las técnicas y metodologías empleadas en la investigación a través de las secciones "Metodología y Técnica", "Aspectos Técnicos y Anexos", "Contexto y Ubicación", y "Claridad y Definición". Además, "Metodología y Datos", "Variables e Instrumentos", y "Variables y Medición" se centraron en la recopilación y análisis de datos, proporcionando una comprensión detallada de la metodología utilizada.

La presentación y redacción adecuadas fueron aspectos cruciales que se resaltaron en las secciones "Redacción y Mejora", "Edición y Formato", "Formato y Presentación", y "Título y Bibliografía". Estas secciones enfatizaron la importancia de comunicar los resultados de la investigación de manera clara y coherente. Además, se hizo hincapié en la precisión de la escritura en las secciones "Redacción y Corrección" y "Redacción y Ortografía".

La organización y planificación del estudio se exploraron en las secciones "Aspectos Organizativos", "Organización y Planificación", y "Justificación y Relevancia", brindando una visión completa sobre cómo se estructuró y planificó la investigación.

Las referencias y normativas pertinentes fueron destacadas en las secciones "Definición y Objetivos", "Normas y Citación", y "Normas APA y Citación", subrayando la importancia de citar adecuadamente las fuentes utilizadas y seguir las normativas específicas.

Finalmente, los resultados, síntesis y modelos derivados de la investigación fueron presentados en las secciones "Modelos y Evaluación", "Síntesis y Recomendaciones", "Resumen y Contenido", y "Modelo y Evaluación", proporcionando una visión integral de los hallazgos, interpretaciones y modelos generados como resultado de la investigación.

Tabla 4

Análisis de Regresión

Medida / Variable	Perplejidad	Tiempo (minutos)
Método	Least Squares	Least Squares
Df Residuales	2	2
Df Modelo	1	1
Coef. (const)	-6.835	3.345
Coef. (Número de Tópicos)	-0.2786	0.0956
Std. Error (Número de Tópicos)	0.033	0.003
t-value (Número de Tópicos)	-8.394	32.374
p-valor (Número de Tópicos)	0.014	0.001
R^2	0.972	0.998
Adj. R^2	0.959	0.997

CONCLUSIONES

Una La aplicación de la modelización LDA demostró ser una técnica efectiva para la identificación de temas relevantes en las revisiones de tesis universitarias. Los resultados obtenidos indicaron que LDA pudo clasificar con precisión los comentarios de revisores en categorías temáticas significativas, lo que facilitó una mejor comprensión de la información contenida en las revisiones.

La implementación de la modelización LDA se llevó a cabo de manera exitosa, lo que permitió la identificación de temas latentes en las revisiones de tesis universitarias. Esto representó un avance significativo en la automatización del proceso de clasificación temática, reduciendo la carga de trabajo asociada a la categorización manual.

La capacidad de la modelización LDA para identificar temas relevantes y recurrentes se confirmó mediante el análisis de los resultados. Los temas identificados por LDA fueron coherentes y se alinearon con los contenidos de las revisiones de tesis, lo que validó su eficacia en la tarea de clasificación temática.

La evaluación de la precisión y coherencia de los temas identificados por LDA arrojó resultados prometedores. La comparación con otros métodos de identificación de temas demostró que LDA proporcionó una mayor precisión en la identificación de temas relevantes y recurrentes, respaldando así la utilidad de esta técnica en la caracterización de revisiones de tesis.

RECOMENDACIONES

Visto el análisis de los resultados y conclusiones la aplicación y uso de la modelización Latent Dirichlet Allocation (LDA) para la identificación y categorización de temas en las revisiones de tesis universitarias resulta una técnica efectiva. Este método ha probado no sólo ser eficaz, sino también eficiente, facilitando la comprensión y el análisis de los comentarios recibidos en dichas revisiones.

Es aconsejable también considerar la implementación de LDA como una herramienta estándar para la automatización del proceso de clasificación temática en futuras revisiones de tesis. Este enfoque no sólo optimizará el tiempo y esfuerzo dedicado a la categorización manual, sino que también promoverá una mayor consistencia y precisión en la identificación de temas relevantes y recurrentes en los textos revisados. La reducción de carga de trabajo en los procesos manuales derivará en un análisis más ágil y profundo de los contenidos.

Adicionalmente, sería provechoso realizar estudios continuos y evaluaciones periódicas del desempeño de la modelización LDA en diferentes contextos y conjuntos de datos. Esto permitiría ajustar y refinar el método para mejorar su eficacia y precisión a lo largo del tiempo, adaptándose a posibles cambios y tendencias en los temas y estilos de escritura presentes en las tesis universitarias.

Es igualmente prudente comparar constantemente los resultados obtenidos mediante LDA con los de otros métodos de identificación de temas. Tal comparación no solo validará continuamente la superioridad y utilidad de LDA, sino que también proporcionará insights valiosos para mejorar y optimizar esta técnica conforme surjan nuevos enfoques y tecnologías en el campo de la identificación y clasificación temática.

BIBLIOGRAFÍA

Alrumiah, S. S., & Al-Shargabi, A. A. (2022). Educational Videos Subtitles' Summarization Using Latent Dirichlet Allocation and Length Enhancement. *Cmc-Computers Materials & Continua*. https://doi.org/10.32604/cmc.2022.021780

Arnautov, S., Trach, B., Gregor, F., Knauth, T., Martin, A., Priebe, C., Lind, J., Muthukumaran, D., O'Keeffe, D., Stillwell, M., Goltzsche, D., Eyers, D., Kapitza, R., Pietzuch, P., & Fetzer, C. (2016). SCONE: secure Linux containers with Intel SGX. *USENIX Symposium on Operating Systems Design and Implementation*. https://www.semanticscholar.org/paper/60dd01cf706c151d2d0b68de27afe22f47e3a9f1

Backenroth, D., He, Z., Kiryluk, K., Boeva, V., Pethukova, L., Khurana, E., Christiano, A. M., Buxbaum, J. D., & Ionita-Laza, I. (2017). FUN-LDA: A latent Dirichlet allocation model for predicting tissue-specific functional effects of noncoding variation. *BioRxiv*. https://doi.org/10.1101/069229

Bastani, K., Namavari, H., & Shaffer, J. (2018). Latent Dirichlet Allocation (LDA) for Topic Modeling of the CFPB Consumer Complaints. *ArXiv: Information Retrieval*. https://doi.org/10.1016/j.eswa.2019.03.001

Billami, M. B., Bortolaso, C., & Derras, M. (2020). Extraction de thèmes d'un corpus de demandes de support pour un logiciel de relation citoyen. *JEPTALNRECITAL*. https://www.semanticscholar.org/paper/370e4acb1a5efa7e67cb7d424787891c6ef02976

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*. https://scholar.google.com/scholar?q=Latent dirichlet allocation

Calvo-Valverde, L. A., Calvo-Valverde, L. A., & Mena-Arias, J. A. (2020). *Evaluacion de distintas tecnicas de representacion de texto y medidas de distancia de texto usando knn para clasificacion de documentos*. https://doi.org/10.18845/tm.v33i1.5022

Cazalens, S., Pacitti, E., Calabretto, S., & Yang, Y. (2013). Sur l'utilisation de LDA en RI Pair à Pair. *INFORSID*.

https://www.semanticscholar.org/paper/d3d9a212ceb96270bf48a32e73184623c570873c

Celard, P., Vieira, A. S., Vieira, A. S., Iglesias, E. L., & Borrajo, L. (2020). LDA filter: A Latent Dirichlet Allocation preprocess method for Weka. *PLOS ONE*. https://doi.org/10.1371/journal.pone.0241701

Chauhan, U., & Shah, A. (2021). Topic Modeling Using Latent Dirichlet allocation: A Survey. *ACM Computing Surveys*. https://doi.org/10.1145/3462478

Chen, Z., Chen, Z., Zhang, Y., Zhang, Y., Wu, C., & Ran, B. (2019). Understanding Individualization Driving States via Latent Dirichlet Allocation Model. *IEEE Intelligent Transportation Systems Magazine*. https://doi.org/10.1109/mits.2019.2903525

Cheng, X., Cheng, X., Cao, Q., & Liao, S. S. (2020). An overview of literature on COVID-19, MERS and SARS: Using text mining and latent Dirichlet allocation. *Journal of Information Science*. https://doi.org/10.1177/0165551520954674

Daud, A., Shamshirband, S., Daud, A., Khan, J. A., Nasir, J. A., Abbasi, R. A., Aljohani, N. R., & Alowibdi, J. S. (2018). Latent Dirichlet Allocation and POS Tags Based Method for External Plagiarism Detection: LDA and POS Tags Based Plagiarism Detection. *International Journal on Semantic Web and Information Systems*. https://doi.org/10.4018/978-1-5225-8057-7.ch015

Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and Effective Latent Concept Modeling for Ad Hoc Information Retrieval. *Document Numérique*. https://doi.org/10.3166/dn.17.1.61-84

Egger, R., & Yu, J. (2022). A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology*. https://doi.org/10.3389/fsoc.2022.886498

Grisales-Aguirre, A. M., & Figueroa-Vallejo, C. J. (2022). Modelado de tópicos aplicado al análisis del papel del aprendizaje automático en revisiones sistemáticas. *Revista de Investigación, Desarrollo e Innovación*. https://doi.org/10.19053/20278306.v12.n2.2022.15271

Gropp, C., Herzog, A., Safro, I., Wilson, P. W., & Apon, A. (2019). Clustered Latent Dirichlet Allocation for Scientific Discovery. *2019 IEEE International Conference on Big Data (Big Data)*. https://doi.org/10.1109/bigdata47090.2019.9005964

Hoblos, J. (2020). Experimenting with Latent Semantic Analysis and Latent Dirichlet Allocation on Automated Essay Grading. *International Conference on Social Networks Analysis, Management and Security*. https://doi.org/10.1109/snams52053.2020.9336533

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., Zhao, L., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-018-6894-4

Kang, H. J., Kang, H. J., Kim, C., Kim, C., Kang, K., & Kang, K. (2019). Analysis of the Trends in Biochemical Research Using Latent Dirichlet Allocation (LDA). *Processes*. https://doi.org/10.3390/pr7060379

Kanungsukkasem, N., & Leelanupab, T. (2019). Financial Latent Dirichlet Allocation (FinLDA): Feature Extraction in Text and Data Mining for Financial Time Series Prediction. *IEEE Access*. https://doi.org/10.1109/access.2019.2919993

Kiatkawsin, K., Sutherland, I., Kim, J.-Y., & Kim, J.-Y. (2020). A Comparative Automated Text Analysis of Airbnb Reviews in Hong Kong and Singapore Using Latent Dirichlet Allocation. *Sustainability*. https://doi.org/10.3390/su12166673

Kozlowski, D., Kozlowski, D., Semeshenko, V., Molinari, A., & Molinari, A. (2020). Latent dirichlet allocation models for world trade analysis. *ArXiv: Physics and Society*. https://doi.org/10.1371/journal.pone.0245393

Li, Z., White, J. C., Wulder, M. A., Coops, N. C., Hermosilla, T., Davidson, A., & Comber, A. (2020). Land cover harmonization using Latent Dirichlet Allocation. *International Journal of Geographical Information Science*. https://doi.org/10.1080/13658816.2020.1796131

Liu, Y., Fei, D., Du, F., Du, F., Sun, J., & Jiang, Y. (2020). ILDA: An interactive latent Dirichlet allocation model to improve topic quality: *Journal of Information Science*. https://doi.org/10.1177/0165551518822455

Martínez-Comeche, J. (2023). Veinticinco años de investigación en redes sociales: evolución de temas entre 1997 y 2021 empleando el algoritmo Asignación Latente de Dirichlet. *Investigación Bibliotecológica: Archivonomía, Bibliotecología e Información*. https://doi.org/10.22201/iibi.24488321xe.2023.96.58777

Morchid, M., & Linar`es, G. (2012). Extraction de mots clefs dans des vid'eos Web par Analyse Latente de Dirichlet (LDA-based tagging of Web videos) [in French]. *JEP/TALN/RECITAL*. https://www.semanticscholar.org/paper/2d65dce38750919008493abdc3d5a6e798127043

Nallapati, R., & Cohen, W. W. (2008). Link-PLSA-LDA: A New Unsupervised Model for Topics and Influence of Blogs. *International Conference on Web and Social Media*. https://doi.org/10.1609/icwsm.v2i1.18621

Negara, E. S., Triadi, D., & Andryani, R. (2019). Topic Modelling Twitter Data with Latent Dirichlet Allocation Method. *2019 International Conference on Electrical Engineering and Computer Science (ICECOS)*. https://doi.org/10.1109/icecos47637.2019.8984523

Nzali, M. D. T., Bringay, S., Lavergne, C., Lavergne, C., Mollevi, C., & Mollevi, C. (2016). *De quoi parlent les patients dans les forums de santé : classification non-supervisée par LDA*. https://www.semanticscholar.org/paper/43bb707083cfd89da2d9a4f71291f53d9b66e162

Osmani, A., Bagherzadeh, J., Mohasefi, J. B., & Gharehchopogh, F. S. (2020). Enriched Latent Dirichlet Allocation for Sentiment Analysis. *Expert Systems*. https://doi.org/10.1111/exsy.12527

Priyantina, R. A., & Sarno, R. (2019). Sentiment Analysis of Hotel Reviews Using Latent Dirichlet Allocation, Semantic Similarity and LSTM. *International Journal of Intelligent Engineering and Systems*. https://doi.org/10.22266/ijies2019.0831.14

Putri, I. R., Kusumaningrum, R., & Kusumaningrum, R. (2017). *Latent Dirichlet Allocation (LDA) for Sentiment Analysis Toward Tourism Review in Indonesia*. https://doi.org/10.1088/1742-6596/801/1/012073

Qomariyah, S., Iriawan, N., & Fithriasari, K. (2019). Topic modeling Twitter data using Latent Dirichlet Allocation and Latent Semantic Analysis. *THE 2ND INTERNATIONAL CONFERENCE ON SCIENCE, MATHEMATICS, ENVIRONMENT, AND EDUCATION*. https://doi.org/10.1063/1.5139825

Rieger, J., Rahnenführer, J., & Jentsch, C. (2020). Improving Latent Dirichlet Allocation: On Reliability of the Novel Method LDAPrototype. *International Conference on Applications of Natural Language to Data Bases*. https://doi.org/10.1007/978-3-030-51310-8_11

Robalino, B., & Stalin, R. (2019). *Método de concordancia bayesiano y su aplicación en problemas de clasificación multiclase con categorías desequilibradas*. https://www.semanticscholar.org/paper/c477b9dbb8d436f5cec93955b98c0997ce52be21

Savoy, J., Savoy, J., & Savoy, J. (2012). *Attribution d'auteur : Une approche basée sur l'allocation latente de Dirichlet (LDA)*. https://www.semanticscholar.org/paper/706aa00b1ca29c1e570463fcd1566ee7f1c2fa5d

Shakeel, K., Tahir, G. R., Tehseen, I., Ali, M., & Ali, M. (2018). A framework of Urdu topic modeling using latent dirichlet allocation (LDA). *Computing and Communication Workshop and Conference*. https://doi.org/10.1109/ccwc.2018.8301655

Suadaa, L. H., Suadaa, L. H., & Purwarianti, A. (2016). Combination of Latent Dirichlet Allocation (LDA) and Term Frequency-Inverse Cluster Frequency (TFxICF) in Indonesian text clustering with labeling. *International Conference on Information and Communicatiaon Technology*. https://doi.org/10.1109/icoict.2016.7571885

Syed, S., & Spruit, M. (2017). Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation. *International Conference on Data Science and Advanced Analytics*. https://doi.org/10.1109/dsaa.2017.61

Toubia, O., Iyengar, G., Iyengar, G., Bunnell, R., & Lemaire, A. (2019). Extracting Features of Entertainment Products: A Guided Latent Dirichlet Allocation

Approach Informed by the Psychology of Media Consumption: *Journal of Marketing Research*. https://doi.org/10.1177/0022243718820559

Tran, B. X., Latkin, C. A., Sharafeldin, N., Nguyen, K., Vu, G. T., Tam, W. W. S., Cheung, N.-M., Nguyen, H. L. T., Ho, C. S. H., & Ho, R. C. M. (2019). Characterizing artificial intelligence applications in cancer research: A latent dirichlet allocation analysis. *JMIR Medical Informatics*. https://doi.org/10.2196/14401

Wahyudi, E., & Kusumaningrum, R. (2019). Aspect Based Sentiment Analysis in E-Commerce User Reviews Using Latent Dirichlet Allocation (LDA) and Sentiment Lexicon. *2019 3rd International Conference on Informatics and Computational Sciences (ICICoS)*. https://doi.org/10.1109/icicos48119.2019.8982522

Wang, R., Hao, J.-X., Law, R., & Wang, J. (2019). Examining destination images from travel blogs: a big data analytical approach using latent Dirichlet allocation. *Asia Pacific Journal of Tourism Research*. https://doi.org/10.1080/10941665.2019.1665558

Wang, Y., & Taylor, J. E. (2019). DUET: Data-Driven Approach Based on Latent Dirichlet Allocation Topic Modeling. *Journal of Computing in Civil Engineering*. https://doi.org/10.1061/(asce)cp.1943-5487.0000819

Wang, Y., Tong, Y., & Shi, D. (2020). Federated Latent Dirichlet Allocation: A Local Differential Privacy Based Framework. *AAAI Conference on Artificial Intelligence*. https://doi.org/10.1609/aaai.v34i04.6096

Webb, J. B., Bray, A., Asare, P., Clipp, R. B., Mehta, Y. B., Mehta, Y. B., Penupolu, S., Penupolu, S., Penupolu, S., Patel, A. A., Poler, S. M., & Poler, S. M. (2020). Computational Simulation to Assess Patient Safety of Uncompensated COVID-19 Two-patient Ventilator Sharing Using the Pulse Physiology Engine. *MedRxiv*. https://doi.org/10.1101/2020.05.19.20107201

Xue, J., Chen, J., Chen, C., Chen, C., Zheng, C., Li, S., Zhu, T., & Zhu, T. (2020). Public discourse and sentiment during the COVID 19 pandemic: Using Latent Dirichlet Allocation for topic modeling on Twitter. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0239441

Ye, T., Li, G., Ahmad, I., Zhang, C., Lin, X., & Li, J. (2021). FLAG: Few-shot Latent Dirichlet Generative Learning for Semantic-aware Traffic Detection. *IEEE Transactions on Network and Service Management*. https://doi.org/10.1109/tnsm.2021.3131266

Zhang, W., Zhang, W., Zhang, W., Zhang, W., Cui, Y., & Yoshida, T. (2017). En-LDA: An Novel Approach to Automatic Bug Report Assignment with Entropy Optimized Latent Dirichlet Allocation. *Entropy*. https://doi.org/10.3390/e19050173

Zhao, F., Ren, X., Yang, S., Han, Q., Zhao, P., & Yang, X. (2020). Latent Dirichlet Allocation Model Training with Differential Privacy. *ArXiv: Learning*. https://doi.org/10.1109/tifs.2020.3032021

Zhou, S., Kan, P., Huang, Q., Huang, Q., Huang, Q., & Silbernagel, J. (2021). A guided latent Dirichlet allocation approach to investigate real-time latent topics of Twitter data during Hurricane Laura: *Journal of Information Science*. https://doi.org/10.1177/01655515211007724

ANEXOS