Universidad de La Habana
Facultad de Matemática y Computación

Análisis de la Evolución de Tópicos en Publicaciones Médicas Usando Modelos de Tópicos Dinámicos

Autor:
Naomi Lahera Champagne

Tutor:
Dr. Luciano García Garrido

Trabajo de Diploma
presentado en opción al título de
Licenciado en Ciencia de la Computación

7 de febrero de 2025

Agradecimientos

Quiero expresar mi más profunda gratitud a mi familia, cuyo apoyo incondicional y amor han sido fundamentales en cada momento, a mis amigos, por su compañía y aliento, y a mi tutor, por su guía y paciencia. A todos ustedes, muchas gracias.

Opinión del tutor

La topicalización de documentos continua recibiendo una atención relevante para la solución de problemas relacionados con su agrupación y clasificación para la rápida identificación de contenidos. La generación de resúmenes entre otros. Las soluciones a su vez, constituyen fundamento esencial de diferentes procesos como la investigación científica, los análisis de la documentación industrial, comercial, etc. El área ha desarrollado la modelación de tres dimensiones relevantes de la topicalización: la detección de tópicos presentes en una colección de documentos, su posible estructuración jerárquica y finalmente su evolución en el tiempo.

El trabajo de diploma de la estudiante Naomi Lahera Champagne desarrolla por primera vez en nuestro grupo de Inteligencia Artificial la tercera de las dimensiones, es decir, la evolución de los tópicos en el tiempo de un corpus de documentos. Para ello evalúa dos de los modelos más recientes de la dinámica de tópicos y como aspecto esencial de sus desarrollos los utiliza para detectar la evolución de los tópicos en un corpus de una revista médica cubana. Los resultados ya se han comenzado a analizar con profesionales del patio y continuarán en el futuro.

La estudiante se dedicó a la tarea planteada con motivación y fue solucionando todos los problemas y sugerencias que le fueron planteadas. En consecuencia, considero que la estudiante Naomi Lahera Champagne ha alcanzado el nivel de profesionalidad que le permite obtener el título de Lic. en Ciencia de la Computación y solicitamos a este tribunal la calificación de Excelente (5) para su trabajo de diploma.

La Habana, 7 de febrero de 2025

Dr. Luciano García Garrido
Profesor Titular Consultante
Facultad de Matemática y Computación
Universidad de La Habana, Cuba

Resumen

Los modelos de tópicos son herramientas esenciales para extraer estructuras temáticas latentes en grandes volúmenes de texto. Los modelos de tópicos dinámicos permiten analizar la evolución de los tópicos en publicaciones científicas a lo largo del tiempo. En este estudio se evaluaron los modelos de tópicos dinámicos *Dynamic Embedded Topic Model (DETM)* y *Chain-Free Dynamic Topic Model* (CFDTM) sobre un conjunto de datos compuesto por artículos de la Revista de Ciencias Médicas de La Habana, Cuba. El objetivo fue identificar cuál modelo se adapta mejor a los datos, utilizando métricas de coherencia de tópicos, diversidad de tópicos, *clustering* y clasificación. Tras una serie de experimentos, se determinaron las configuraciones óptimas de hiperparámetros para ambos modelos y se comparó su desempeño. Los resultados revelan que, aunque ambos modelos presentan una coherencia similar, CFDTM supera notablemente a DETM en términos de diversidad de tópicos y agrupación de documentos (*clustering*), lo que indica una mejor captura de la evolución temporal de los tópicos en la investigación médica.

Abstract

Topic models are essential tools for extracting latent thematic structures from large text corpora. Dynamic topic models allow the analysis of topic evolution in scientific publications over time. This study evaluates the dynamic topic models *Dynamic Embedded Topic Model (DETM)* and *Chain-Free Dynamic Topic Model (CFDTM)* on a dataset composed of articles from the *Revista de Ciencias Médicas de La Habana*, Cuba. The objective was to identify which model best fits the data, using metrics such as topic coherence, topic diversity, clustering, and classification. After a series of experiments, the optimal hyperparameter configurations for both models were determined, and their performance was compared. The results reveal that, although both models exhibit similar coherence, *CFDTM* significantly outperforms *DETM* in terms of topic diversity and document clustering, indicating a better capture of the temporal evolution of topics in medical research.

Índice general

Índice de figuras

Índice de tablas

Introducción

Los modelos de tópicos son herramientas matemáticas-computacionales que se diseñan para extraer la estructura temática latente en una colección de documentos. Esos modelos facilitan la automatización de tareas como el análisis semántico, la recomendación de contenido y la identificación de relaciones lingüísticas, que incluye fenómenos como la sinonimia y la polisemia. Además, su versatilidad permite aplicarlos en diversas áreas, como el análisis de opiniones, la categorización automática de textos y la recuperación de información.

Un modelo de tópicos analiza un conjunto de documentos, a lo que se llama corpus, dentro del cual se identifican todas las palabras únicas que forman el vocabulario. El objetivo de los modelos de tópicos es descubrir los tópicos subyacentes en esos documentos. Cada tópico se define por un conjunto de palabras del vocabulario y, a su vez, cada documento se puede asociar a una combinación de estos tópicos.

Existen diversos enfoques, como los modelos estáticos que ofrecen una visión instantánea de los tópicos, y los modelos jerárquicos que establecen relaciones de subordinación y especialización entre tópicos lo que permite reflejar como los tópicos más generales engloban a los más específicos. Sin embargo, para analizar cómo evolucionan los tópicos a lo largo del tiempo existen los modelos de tópicos dinámicos. A diferencia de los enfoques estáticos, esos modelos rastrean cómo los tópicos surgen, se modifican o se desvanecen a los largo del tiempo, lo que modela la evolución de su contenido.

El campo de los modelos de tópicos es objeto de un estudio exhaustivo a nivel mundial, y Cuba no es una excepción. En el Grupo de Inteligencia Artificial de la Universidad de La Habana, Cuba, se han desarrollado modelos estáticos y jerárquicos para la extracción de tópicos en diversos corpus de texto. El problema es que, a pesar de contar con modelos de tópicos eficientes en Cuba, no es posible comprender cómo evolucionan los tópicos de investigación en revistas científicas como la Revista de Ciencias Médicas de La Habana. Esa limitación dificulta la identificación de tendencias y la toma de decisiones informadas sobre la investigación.

Dado el interés en comprender cómo evolucionan los tópicos de investigación

en la Revista de Ciencias Médicas de La Habana, Cuba, resulta esencial utilizar modelos que puedan capturar la dimensión temporal de los datos. Los modelos de tópicos dinámicos son especialmente adecuados para este propósito, ya que permiten identificar tendencias y patrones a lo largo del tiempo.

A partir de esa situación la presente investigación pretende responder la siguiente pregunta científica ¿Cuál de los modelos de tópicos dinámicos mejora el desempeño para indagar en la evolución de los tópicos subyacentes en la Revista de Ciencias Médicas de La Habana?

Existen diversos modelos de tópicos dinámicos que permiten capturar la evolución temática en conjuntos de datos textuales, cada uno con características particulares que los hacen adecuados para diferentes contextos. Sin embargo, destacan en especial los modelos *Dynamic Embedded Topic Model* (DETM) (Adji B. Dieng 2019) y *Chain-Free Dynamic Topic Model* (CFDTM) (Wu, Dong et al. 2024) por sus enfoques avanzados y complementarios. Estos modelos representan paradigmas contrastantes en cómo abordan la evolución temporal y semántica de los tópicos, lo que los convierte en opciones ideales para un análisis más detallado.

Por un lado, el DETM utiliza cadenas de Markov para modelar cambios graduales y asegurar la continuidad temática a lo largo del tiempo, resultando especialmente útil en datos donde los tópicos evolucionan de forma gradual como las de ciertas áreas científicas. En cambio, el CFDTM, al eliminar la dependencia de las cadenas de Markov, permite una evolución más flexible y adaptativa de los tópicos, siendo más adecuado para datos con cambios abruptos y emergentes. Este contraste asegura que ambos modelos aborden diferentes aspectos de la dinámica temporal, proporcionando una evaluación más completa de las características del conjunto de datos. Además de estas diferencias en su enfoque temporal, ambos modelos ofrecen representaciones semánticas sofisticadas. El DETM emplea *embeddings* para capturar relaciones semánticas complejas entre palabras y tópicos, mientras que el CFDTM introduce innovaciones como *Evolution-Tracking Contrastive Learning* (ETC) y *Unassociated Word Exclusion* (UWE) para garantizar la diversidad entre los tópicos y la correspondencia de las palabras de un tópico con el momento en el tiempo en que fue extraído.

Para dar salida a la pregunta científica se define el objetivo general: determinar cuál de los modelos de tópicos dinámicos, *Dynamic Embedded Topic Model* (DETM) (Adji B. Dieng 2019) o *Chain-Free Dynamic Topic Model* (CFDTM) (Wu, Dong et al. 2024), ofrece el mejor desempeño para rastrear la evolución de tópicos a lo largo del tiempo en la Revista de Ciencias Médicas de La Habana, Cuba.

En consecuencia, para alcanzar este objetivo general, se plantean los siguientes objetivos específicos: en primer lugar, evaluar el rendimiento de los modelos DETM y

CFDTM con diferentes configuraciones de hiperparámetros; en segundo lugar, comparar la efectividad de ambos modelos al utilizar métricas como *Topic Coherence* y *Topic Diversity*, que miden la calidad y diversidad de los tópicos identificados; posteriormente (tercero), proponer un modelo adecuado para el análisis dinámico de tópicos en el contexto de revistas científicas cubanas, específicamente en la Revista de Ciencias Médicas de La Habana para luego analizar la evolución de los tópicos relevantes en la misma, demostrando la aplicabilidad del modelo seleccionado.

Esta tesis, desarrollada en el Grupo de Inteligencia Artificial de la Facultad de Matemática y Computación de la Universidad de La Habana, Cuba, se estructura en tres capítulos principales: Estado del Arte (Capítulo 1), Detalles de Implementación (Capítulo 2) y Experimentos y Resultados (Capítulo 3). El primer capítulo presenta una revisión exhaustiva de los trabajos previos relacionados con los modelos de tópicos dinámicos, mientras que los capítulos subsiguientes detallan el proceso de implementación, los experimentos realizados y los hallazgos obtenidos.

Capítulo 1

Estado del Arte

Los modelos de tópicos, diseñados para descubrir los tópicos latentes en colecciones de documentos, evolucionaron significativamente en las últimas décadas. Desde los enfoques bayesianos tradicionales hasta los modelos neuronales más recientes.

Los enfoques clásicos para modelos de tópicos incluyen modelos basados en factorización de matrices no negativas, como el *Latent Semantic Analysis* (LSA) (Deerwester et al. 1988), que inicialmente define una matriz término-documento y luego aplica descomposición en valores singulares (SVD) a dicha matriz. Mediante la descomposición las matrices resultantes capturan la relevancia de las palabras en los tópicos, los valores singulares y la representación de los documentos en el espacio temático. Al truncar en función de un K (hiperparámetro) la matriz de valores singulares y reajustar las dimensiones de las matrices restantes se obtiene una representación reducida que ofrece proporciones de tópicos por documento y palabras por tópico. LSA fue ampliamente utilizado en recuperación de información. Sin embargo, presenta limitaciones como la selección del parámetro K, el alto consumo de recursos computacionales y la dificultad de interpretar valores negativos en sus resultados.

Para superar estas deficiencias surgió el modelo *Probabilistic Latent Semantic Analysis* (PLSA) (Hofmann 1999), un modelo probabilístico que asume que cada documento es una distribución multinomial de tópicos latentes. PLSA utiliza una red bayesiana para modelar las relaciones entre las variables (tópicos como latentes y palabras/documentos como observables) siguiendo un enfoque probabilístico basado en modelos de clases latentes y para optimizar los parámetros aplica el algoritmo *Expectation-Maximization* (EM) (Arthur P. Dempster 1977).

El modelo *Latent Dirichlet Allocation* (LDA) (D. M. Blei, Ng y Jordan 2003) es otro enfoque probabilístico que define un tópico como una distribución sobre un vocabulario predefinido. Ese modelo considera que varios tópicos pueden ocurrir de

forma simultánea en un documento, compartiendo todos los documentos los mismos K tópicos con proporciones distintas.

El proceso de LDA consta de tres fases principales: inicialización, reasignación y optimización. Al comenzar con la inicialización, cada palabra se asigna de forma independiente a un tópico. Luego durante la reasignación, utilizando distribuciones multinomiales generadas por Dirichlet, se ajustan las asignaciones previas considerando la frecuencia de uso de cada tópico en un documento y la frecuencia con la que un tópico contiene ciertas palabras. Finalmente, en la optimización, las reasignaciones iterativas se refinan hasta alcanzar un criterio de convergencia definido por la mínima variación en las asignaciones o un número fijo de iteraciones.

Latente Dirichlet Allocation (LDA) sentó las bases para modelos que exploran tanto la estructura jerárquica como la evolución temporal de los tópicos. Modelos como *Raphil* (Murphy 2012) y *Pachinko Allocation Model* (Li y McCallum 2006) representan tópicos en forma de árbol, mientras que *Topic Over Time* (Wang y McCallum 2006) captura como estos tópicos evolucionan a lo largo del tiempo. Estos modelos permiten realizar análisis más detallados y sofisticados de grandes volúmenes de texto.

Si bien LDA es preciso y ampliamente usado, presenta limitaciones al trabajar con grandes volúmenes de datos debido al aumento de variables según el tamaño del corpus y la cantidad de tópicos. Las limitaciones de los modelos de tópicos tradicionales impulsaron la investigación hacia enfoques más sofisticados. Los modelos de tópicos neuronales, basados en redes neuronales profundas, surgieron como una respuesta a estas limitaciones. Estos modelos aprovechan el poder computacional de las redes neuronales para inferir los tópicos latentes en un corpus de documentos de manera más eficiente y precisa.

Los modelos de tópicos neuronales han incorporado diversas técnicas para mejorar la calidad de la inferencia y la representatividad de los tópicos extraídos. Los modelos basados en autoencoders (AE-based NTMs) utilizan redes neuronales para mapear documentos a una representación latente y aprender distribuciones de tópicos de manera eficiente. Los modelos de tópicos con *embeddings* (NTMs with Embeddings) aprovechan representaciones vectoriales de palabras y documentos para mejorar la coherencia semántica de los tópicos aprendidos. Por otro lado, los modelos de tópicos con entrenamiento adversarial (NTMs with Adversarial Training) emplean redes generativas y discriminativas para mejorar la separación y estabilidad de los tópicos inferidos. Además, los modelos de tópicos con metadatos (NTMs with Metadata) incorporan información adicional, como autores o etiquetas, para guiar la distribución de los tópicos de manera más informada. En este contexto, los modelos de tópicos dinámicos extienden estas ideas al dominio temporal, permitiendo capturar la evolu-

ción de los tópicos a lo largo del tiempo y proporcionando una visión más completa de la variabilidad en colecciones de documentos.

Los Modelos de Tópicos Neuronales Dinámicos, DNTM por sus siglas en inglés, evolucionaron principalmente en la última década. Las variantes propuestas alcanzaron gran precisión y entre ellas difieren incluso en la modelación del problema. A diferencia del enfoque tradicional donde los documentos están representados por una bolsa de palabras [1] y los tópicos están representados por un vector que indica la proporción de las palabras en dicho tópico, surgieron modelos como *Dynamic Embedded Topic Model (D-ETM)* (Adji B. Dieng 2019) que sigue un enfoque diferente representando palabras y tópicos como *embeddings* [2]. Este modelo tiene como bases a los modelos LDA (D. M. Blei, Ng y Jordan 2003), *Dynamic LDA* (D-LDA) (Blei y Lafferty 2006) y *Embedded Topic Model* (ETM) (Dieng et al. 2019). D-LDA es una extensión de LDA que utiliza una serie de tiempo probabilística para permitir que los tópicos varíen suavemente a lo largo del tiempo, pero disminuye su factibilidad al presentar las misma limitaciones que LDA. Para sobrepasar esas limitaciones surge ETM que representa los tópicos como un vector en el espacio de los *embeddings* de las palabras para luego usar el producto punto entre los *embeddings* de las palabras y los *embeddings* de los tópicos para definir la distribución de palabras por tópico. De manera similar a D-LDA, el D-ETM involucra una serie de tiempo probabilística para permitir que los tópicos varíen suavemente a lo largo del tiempo. Sin embargo, cada tópico en el D-ETM es un vector que varía con el tiempo en el espacio en el que fue proyectado. Al igual que en el ETM, la probabilidad de cada palabra bajo el D-ETM es una distribución categórica cuyo parámetro natural depende del producto interno entre el *embedding* de la palabra y el *embedding* de su tópico asignado. En contraste con el ETM, los *embedding* de tópico del D-ETM varían con el tiempo y son aparoximados mediante el mecanismo de inferencia variacional (Jordan et al. 1999; D. M. Blei, Kucukelbir y McAuliffe 2017). Para escalar el algoritmo a grandes conjuntos de datos en ese modelo fueron utilizados el submuestreo de datos (Hoffman et al. 2013) y la amortización (Gershman y Goodman 2014). Además, fue utilizada una aproximación variacional estructurada parametrizada por una red de memoria a corto y largo plazo (LSTM) (Hochreiter y Schmidhuber 1997).

Surgieron luego, en 2022 los modelso NetDTM y NetDTM++ (Zhang y Lauw 2022, 17–23 Jul) que funcionan bajo el supuesto de que la naturaleza temporal se relaciona no solo con el momento en que se crea un documento, sino también con

[1] En procesamiento de lenguaje natural, la bolsa de palabras es una representación simplificada de texto donde un documento se describe mediante la frecuencia de aparición de cada palabra, ignorando el orden y la gramática.

[2] Un *embedding* es el mapeo de una variable categórica a un vector numérico.

como los documentos creados en diferentes momentos pueden formar vínculos. Dado que una red de documentos no emerge en su totalidad de forma imprevista, los creadores afirman que teniendo una red inicial de documentos a medida que pasa el tiempo crece no solo el corpus sino también la conectividad de la red. Estos modelos aprovechan esta estructura en red al incorporar las conexiones entre documentos como una fuente adicional de información, mejorando así la capacidad del modelo para capturar patrones temporales y relacionales en los datos.

Otro enfoque novedoso fue el usado en el modelo *Aligned Neural Topic Model* (ANTM) (Rahimi et al. 2023) que descubre la evolución de los tópicos usando *clusters*. ANTM descubre tópicos en evolución a través de un algoritmo de ventanas deslizantes superpuestas para la agrupación temporal de documentos. El enfoque utilizado, llamado *Aligned Clustering*, consiste en segmentar archivos en segmentos superpuestos, realizar *clustering* secuencial basado en densidad y alinear *clusters* de documentos similares adyacentes a lo largo de diferentes períodos. El *Aligned Clustering* permite la identificación de conjuntos de tópicos similares que abarcan múltiples períodos de tiempo y, sin embargo, son lo suficientemente diferentes como para mostrar algún tipo de evolución. ANTM aprovecha los grandes modelos de lenguaje (LLMs) preentrenados, como Data2Vec (Baevski et al. 2022), que predice representaciones latentes de los documentos de manera auto-supervisada utilizando una arquitectura *transformer* estándar.

Siguiendo la idea de establecer conexiones entre los documentos del corpus ubicados en distintos espacios de tiempo surge el modelo *Dynamic Structured Neural Topic Model* (DSNTM) (Miyamoto et al. 2023, julio), un modelo que calcula las dependencias entre los tópicos a través del tiempo. DSNTM modela dichas dependencias basándose en un mecanismo de autoatención (Vaswani et al. 2023; Lin et al. 2017), que al observar sus pesos revela como los tópicos pasados se ramifican o fusionan en nuevos tópicos siendo posible predecir tópicos emergentes y evaluar cuantitativamente qué tópicos pasados contribuyen a la aparición de nuevos tópicos . Este modelo introduce la regularización de citas que lleva a los pesos del mecanismo de atención a reflejar las relaciones de citación entre documentos y como resultado es posible modelar tanto el texto como las citas de forma conjunta, mejorando la calidad de los tópicos inferidos y capturando con precisión sus transiciones. Debido al alto poder expresivo del mecanismo de autoatención y la información adicional de citas DSNTM resalta en factibilidad a la hora de detectar complejos procesos de ramificación y fusión de tópicos a lo largo del tiempo.

En ocasiones categorizar un modelo de tópicos puede ser algo ambiguo. Un ejemplo de esto es el modelo *Neural Dynamic Focused Topic Model* (NDF-TM) (Cvejoski et al. 2023), un modelo que en lugar de modelar la evolución de los tópicos modela

de forma independiente la proporción de tópicos y la ocurrencia de los mismo en los documentos a lo largo del tiempo. Cabe destacar que la ocurrencia de los tópicos evolucionan con el tiempo, pero sus tópicos por sí mismos son invariantes. Por lo tanto, este método no se adhiere con precisión a la definición original de un modelo de tópicos dinámico (Wu, Nguyen y Luu 2024). Este enfoque busca separar la presencia de un tópico en un documento de su proporción dentro del mismo. El resultado es un modelo escalable que permite que el número instantáneo de tópicos activos por documento fluctúe y que desacopla explícitamente la proporción de tópicos de su ocurrencia en los documentos, ofreciendo así capas novedosas de interpretabilidad y transparencia en la evolución de los tópicos a lo largo del tiempo.

En su mayoría los modelos existentes hasta el momento relacionaban los tópicos a través de cadenas de Markov para capturar su evolución y disminuían su efectividad al extraer tópicos repetidos [3] y ubicar tópicos en espacios de tiempo incorrectos.

Según se plantea el artículo *Modeling Dynamic Topics in Chain-Free Fashion by Evolution-Tracking Contrastive Learning and Unassociated Word Exclusion* (Wu, Dong et al. 2024), el uso de cadenas de Markov para encadenar tópicos tiende a agruparlos, lo que dificulta su diferenciación y limita su capacidad para capturar plenamente la semántica de sus respectivos períodos de tiempo. Además, este enfoque puede forzar una relación excesiva entre los tópicos a lo largo de diferentes períodos, reduciendo su asociación con los segmentos de tiempo específicos y ocultando los tópicos genuinos en esos segmentos. Para resolver los problemas anteriores surge el modelo *Chain-Free Dynamic Topic Model* (CFDTM) (Wu, Dong et al. 2024) que rompe la tradición de encadenar tópicos a través de cadenas de Markov y propone un nuevo enfoque respaldado por dos submodelos: *Evolution-Tracking Contrastive Learning* (ETC) y *Unassociated Word Exclusion* (UWE). ETC construye adaptativamente relaciones positivas y negativas entre los tópicos dinámicos. En este contexto ${\varphi_k}^{(t)}$ representa al tópico k en el momento t. Para rastrear la evolución de los tópicos ETC construye relaciones positivas para los pares $({\varphi_k}^{(t-1)}, {\varphi_k}^{(t)})$ y más importante aún, construye relaciones negativas entre los pares $({\varphi_k}^{(t)}, {\varphi_{k'}}^{(t)})$ con $k \neq k'$. Estas relaciones negativas alientan a los tópicos dentro de un mismo espacio de tiempo a ser distintos, manteniendo la diversidad de tópicos y así aliviando el problema de los tópicos repetitivos. Por otro lado UWE encuentra las palabras más relevantes de los tópicos en cada segmento de tiempo e identifica cuáles de ellas no pertenecen a esta segmento de tiempo como palabras no asociadas, así excluye explícitamente estas palabras no asociadas de los tópicos para refinar la semántica de los mismos. Siguiendo la práctica común (Blei y Lafferty 2006; Adji B. Dieng 2019; Miyamoto

[3] Los tópicos dentro de un segmento de tiempo se consediran repetitivos al presentar una semántica similar.

et al. 2023, julio), el modelo usa el hiperparámetro $\lambda(t)$ para ajustar adaptativamente las intensidades de evolución entre segmentos de tiempo. Si los tópicos evolucionan ligeramente entre los segmentos de tiempo $t-1$ y t, se utiliza un $\lambda(t)$ grande; de lo contrario, un $\lambda(t)$ pequeño si evolucionan dramáticamente. El proceso generativo en CFDTM usa un VAE donde el *encoder* es reutilizado para documentos en diferentes espacios de tiempo para garantizar la eficiencia de parámetros. Como resultado de los métodos novedosos empleados, CFDTM alcanzó el mejor índice de TC [4] con mejoras significantes repecto al resto de los modelos existentes.

En conclusión, los Modelos de Tópicos Neuronales Dinámicos han evolucionado considerablemente, incorporando técnicas avanzadas como *embeddings* y mecanismos de autoatención para superar las limitaciones de los enfoques tradicionales. Las innovaciones recientes han enriquecido la capacidad de capturar la evolución semántica y temporal de los tópicos. Estos avances no solo mejoran la interpretabilidad y la diversidad de los tópicos inferidos, sino que también abren nuevas posibilidades para abordar desafíos complejos en el análisis de datos textuales dinámicos.

Con el objetivo de comparar el desempeño de los modelos DETM y CFDTM usando como corpus los artículos publicados en la Revista de Ciencias Médicas de La Habana, Cuba, se realizaron una serie de experimentos en los cuales se determinaron las configuraciones óptimas de hiperparámetros para ambos modelos sobre dicho corpus. En los próximos capítulos se presentan los detalles de este proceso y los resultados obtenidos.

[4] TC: Topic Coherence. Métrica usada para medir la coherencia entre las palabras asignadas a un tópico por un modelo de tópicos.

Capítulo 2

Detalles de Implementación

2.1. *Dataset* y preprocesamiento de los datos

El conjunto de datos (*dataset*) empleado en esta investigación fue compilado a partir de los artículos pertenecientes a la Revista de Ciencias Médicas de La Habana, Cuba. Dicha revista científica cuenta con una colección de más de 1500 artículos disponibles en su sitio web oficial, los cuales constituyeron la fuente para la extracción del *dataset*.

Una vez adquirido el corpus, los documentos fueron sometidos a una fase de preprocesamiento que sigue la metodología implementada en el módulo TopMost (Wu, Pan y Luu 2023) del lenguaje Python, la cual abarca una serie de etapas cruciales para preparar el texto para el modelado de tópicos. Para comenzar, se lleva a cabo la limpieza del texto, que incluye la eliminación de etiquetas HTML, el cambio a minúsculas, la supresión de direcciones de correo electrónico y menciones a usuarios, y además la normalización de espacios en blanco.

Posteriormente, se realiza la tokenización, donde el texto se segmenta en unidades individuales o *tokens*. A continuación, se filtran estos *tokens*, eliminando las palabras de alta frecuencia pero de bajo contenido informativo (*stopwords*) y aquellos *tokens* que no cumplen con ciertos criterios, como la presencia exclusiva de números o una combinación de letras y números. Una vez obtenidos los *tokens* relevantes, se construye el vocabulario, formado por el conjunto de todos los *tokens* presentes en el corpus.

Finalmente, los documentos son representados como bolsas de palabras *(Bag-of-Words)*, donde cada documento se convierte en un vector que indica la frecuencia de cada término del vocabulario en dicho documento. Para enriquecer la representación semántica de los términos, se generan *embeddings*.

2.2. Exploración en el espacio de búsqueda de los hipeparámetros

La eficiencia de los modelos de tópicos dinámicos depende en gran medida del *dataset* empleado, en este caso los artículos pertenecientes a la Revista de Ciencias Médicas de La Habana, Cuba, ya que las características específicas del corpus influyen directamente en la capacidad del modelo para descubrir patrones significativos. Se llevó a cabo una exploración exhaustiva de las configuraciones de hiperparámetros para dos modelos: CFDTM y DETM, con el objetivo de identificar aquellos valores que maximizan la coherencia de los tópicos, la diversidad y el rendimiento general en este corpus.

2.2.1. Aspectos clave del modelo DETM

DETM aborda la evolución temporal modelando los tópicos como secuencias de *embeddings* en un espacio continuo. Además de la distribución de palabras por tópicos, el modelo construye las proporciones de tópicos por documentos (θ). La distribución sobre las proporciones de tópicos $q(\theta_d|\eta_{t_d}, w_d)$ es una distribución logístico-normal, cuyos parámetros de media y covarianza son funciones tanto de la media latente η_{t_d} como de la representación en bolsa de palabras del documento. En particular, estas funciones están parametrizadas por redes neuronales *feed-forward* que toman como entrada η_{t_d} y la representación normalizada en bolsa de palabras.

La distribución sobre las medias latentes $q(\eta_t|\eta_{1:t-1}, w)$ depende de todas las medias latentes previas $\eta_{1:t-1}$. Para capturar esa dependencia temporal, se utiliza una red LSTM. En este modelo se escogió una distribución gaussiana $q(\eta_t|\eta_{1:t-1}, w)$, cuya media y covarianza son determinadas por la salida de la LSTM. La entrada de la LSTM en el tiempo t es el promedio de la representación en bolsa de palabras de todos los documentos con marca de tiempo t. Aquí, w denota la representación normalizada en bolsa de palabras de todos esos documentos.

Notación:

- θ: Representa las proporciones de tópicos en los documentos.
- ρ: Denota las distribuciones de palabras en el espacio latente.
- η: Representa la media utilizada para aproximar θ.

2.2.2. Descripción de hiperparámetros del modelo DETM

- **Número de tópicos** (*num_topics*): Referente al número de tópicos latentes que el modelo intentará descubrir en el corpus. Un valor muy bajo puede mezclar tópicos distintos mientras que un valor muy alto puede granular un mismo tópico.
- **Dimensión de** ρ (*rho_size*) Dimensión de ρ, que representa las distribuciones de palabras en el espacio latente. Puede verse como la dimensión de los tópicos en el espacio de *embeddings*.
- **Tamaño del espacio oculto de** $q(\theta)$ (*en_units*): Tamaño de la capa oculta de la red que aproxima θ ($q(\theta)$), la proporción de los tópicos para cada documento.
- **Función de activación para** $q(\theta)$ (*theta_act*): Especifica la función de activación usada en la red neuronal que aproxima θ ($(q(\theta))$), que representa la proporción de los tópicos para cada documento.
- **Número de capas para** η (*eta_nlayers*): Número de capas de la red neuronal recurrente (RNN) que se usa para modelar la media necesaria para aproximar θ (η).
- **Tamaño del espacio oculto en** η (*eta_hidden_size*): Tamaño de la capa oculta en la RNN que modela la media utilizada para aproximar θ (η).
- δ (*delta*): Varianza del prior en el proceso de evolución temporal de los tópicos en el espacio de *embeddings*. Controla la magnitud del cambio en los tópicos a lo largo del tiempo.
- **Tasa de Aprendizaje** (*learning_rate*): Controla el tamaño de los pasos que da el optimizador durante el entrenamiento. Una tasa de aprendizaje alta puede llevar a la inestabilidad, mientras que una muy baja puede resultar en una convergencia lenta.
- ***Dropout* en el encoder** (*enc_drop*): Define la tasa de *dropout* aplicada a la red neuronal que codifica la distribución $q(\theta)$, usada para estimar la distribución de tópicos por documento.
- ***Dropout* en la RNN para** η (*eta_dropout*): Define la tasa de *dropout* aplicada en la RNN que modela la evolución temporal de los tópicos (η).
- **Número de épocas** (*epochs*): Define el número de veces que el modelo recorrerá el conjunto de datos de entrenamiento completo.

El espacio de búsqueda para encontrar la configuración óptima de hiperparámetros se muestra en las tablas 2.1 y 2.2.

Tabla 2.1: Espacio de búsqueda de hiperparámetros para el modelo DETM

Hiperparámetro	**Valores**
num_topics	{20, 50, 100}
rho_size	{200, 300, 400}
en_units	{400, 800, 1200}
eta_nlayers	{2, 3, 4}
delta	{0.001, 0.005, 0.01, 0.05}
theta_act	{anh, softplus, relu, rrelu, leakyrelu, elu, selu, glu}
eta_hidden_size	{100, 200, 300}

Tabla 2.2: Espacio de búsqueda de hiperparámetros para el entrenamiento del modelo DETM

Hiperparámetro	**Valores**
enc_drop	{0.0, 0.2, 0.3, 0.5}
eta_dropout	{0.0, 0.2, 0.3, 0.5}
learning_rate	{1e-3, 5e-4, 0.02}
epochs	{200, 400, 800}

2.2.3. Aspectos claves del modelo CFDTM

El modelo CFDTM utiliza una arquitectura basada en un *Variational Autoencoder* (VAE) para modelar la distribución de los tópicos en los documentos. Su *encoder* es una red neuronal de dos capas que toma como entrada la representación en bolsa de palabras (BoW) de los documentos y genera como salida los parámetros que representan la distribución latente del documento en el espacio de tópicos. El *decoder* genera documentos a partir de la distribución de tópicos aprendida para calcular la probabilidad de palabras dentro de cada tópico. Además, los *embeddings* de palabras y tópicos se proyectan en un espacio latente de dimensión D utilizando *embeddings* preentrenados y la distribución de palabras en los tópicos se calcula en función de la distancia euclidiana normalizada en este espacio.

Aprendizaje Contrastivo de Seguimiento de Evolución (Evolution-Tracking Contrastive Learning, ETC)

ETC es un mecanismo diseñado para modelar la evolución de los tópicos a lo largo del tiempo y evitar la repetición de tópicos similares dentro de un mismo período. Para lograrlo, se utilizan *embeddings* de los tópicos en un espacio latente. El

aprendizaje contrastivo se implementa mediante la construcción de relaciones entre los tópicos de diferentes períodos y dentro de un mismo período.

Para modelar la evolución de los tópicos, ETC establece relaciones positivas entre los tópicos de distintos períodos de tiempo, acercando los *embeddings* de un mismo tópico en diferentes momentos. Además, para fomentar la diversidad temática y evitar la repetición de tópicos similares dentro de un mismo período, se introducen relaciones negativas que alejan los *embeddings* de tópicos distintos en el mismo período.

Exclusión de Palabras No Asociadas (Unassociated Word Exclusion, UWE)

UWE es un mecanismo diseñado para eliminar palabras que aparecen en tópicos de períodos donde no deberían estar, evitando asociaciones incorrectas. Su implementación se divide en dos pasos: primero, la identificación de palabras no asociadas, y luego su exclusión de los tópicos.

En la identificación, se extraen las palabras más relevantes de cada tópico y se comparan con el vocabulario del período correspondiente. Las palabras que no forman parte del vocabulario de ese período se consideran no asociadas.

Para excluir estas palabras, se ajustan los *embeddings* de los tópicos, empujando sus representaciones lejos de las palabras no asociadas. De este modo, UWE refina la representación de los tópicos y mejora su coherencia con el período al que pertenecen, asegurando que los tópicos reflejen de manera más precisa el contexto temporal de los documentos.

2.2.4. Descripción de hiperparámetros del modelo CFDTM

- **Número de tópicos** (*num_topics*): Referente al número de tópicos latentes que el modelo intentará descubrir en el corpus. Un valor muy bajo puede mezclar tópicos distintos, mientras que un valor muy alto puede granular un mismo tópico.
- **Unidades en la capa encoder** (*en1_units*): Número de neuronas en la primera capa del encoder.
- ***Dropout* en la red** (*dropout*): Define la tasa de *dropout* aplicada para regularizar el modelo.
- **Temperatura de β** (*beta_temp*): Factor de escala para la distribución β de los tópicos en el modelo.
- **Temperatura en contraste** (*temperature*): Controla la suavidad en la función de contraste utilizada en el aprendizaje de *embeddings*.

- **Peso para palabras negativas** (*weight_neg*): Ponderación para las palabras negativas en el mecanismo de contraste.
- **Peso para palabras positivas** (*weight_pos*): Ponderación para palabras positivas en el mecanismo de contraste.
- **Peso para *Unassociated Word Exclusion* (UWE)** (*weight_UWE*): Controla la penalización por palabras que no están asociadas a ningún tópico en el momento t.
- **Número de palabras negativas K** (*neg_topk*): Define cuántas palabras negativas se consideran en la pérdida contrastiva.
- **Tasa de aprendizaje** (*learning_rate*): Controla la magnitud de los cambios en los parámetros del modelo durante el entrenamiento.
- **Tamaño del *batch*** (*batch_size*): Define cuántos ejemplos de entrenamiento se procesan simultáneamente antes de actualizar los parámetros del modelo.
- **Número de épocas** (*num_epoch*): Cantidad de veces que el modelo recorrerá el conjunto de datos completo.

El espacio de búsqueda para encontrar la configuración óptima de hiperparámetros se muestra en las tablas 2.3 y 2.4.

Tabla 2.3: Espacio de búsqueda de hiperparámetros para el modelo CFDTM

Hiperparámetro	**Valores**
num_topics	{20, 50, 100}
en1_units	{50, 100, 200, 300}
dropout	{0.0, 0.1, 0.3, 0.5}
beta_temp	{0.5, 1.0, 1.5, 2.0}
temperature	{0.05, 0.1, 0.2, 0.5}
weight_neg	{1e6, 5e6, 1e7, 1e8}
weight_pos	{1.0, 10.0, 100.0}
weight_UWE	{1e2, 1e3, 1e4}
neg_topk	{5, 10, 15, 20, 30}

2.3. Métricas utilizadas

Para evaluar la calidad del modelo de tópicos, se utilizaron principalmente dos métricas: Coherencia del Tópico (TC - *Topic Coherence*) y Diversidad del Tópico (TD - *Topic Diversity*).

Tabla 2.4: Espacio de búsqueda de hiperparámetros para el entrenamiento del modelo CFDTM

Hiperparámetro	**Valores**
learning_rate	{0.001, 0.002, 0.005, 0.01}
batch_size	{100, 200, 300, 500}
num_epoch	{400, 600, 800, 1000}

Coherencia del Tópico (TC)

La coherencia del tópico mide qué tan relacionadas están las palabras más representativas de un tópico. Esta métrica evalúa si las palabras principales de un tópico aparecen juntas con frecuencia en el corpus de documentos. Para evaluar esta coherencia, se utilizó la métrica CV (Röder et al. 2015). Una mayor coherencia indica que el tópico es más interpretable y significativo.

Diversidad del Tópico (TD)

La diversidad del tópico mide qué tan distintos son los tópicos descubiertos. Se calcula como la proporción de palabras principales de un tópico que no se repiten en otros tópicos. Un alto valor de TD indica que los tópicos son bien diferenciados entre sí, evitando redundancias y asegurando que el modelo capture una variedad más amplia de tópicos.

Además de estas métricas principales, se evaluó la calidad de las distribuciones de tópicos en los documentos mediante dos tareas extrínsecas: clasificación de textos y *clustering* de documentos.

Clasificación de Textos

Para la clasificación de textos se entrena un clasificador tradicional (SVM), utilizando las distribuciones documento-tópico como características. Luego, se predicen las etiquetas de otros documentos, permitiendo evaluar qué tan bien el modelo organiza los documentos en función de sus tópicos.

Clustering de Documentos

Para el *clustering* en este caso, se asigna a cada documento el tópico más significativo en su distribución documento-tópico y se analiza qué tan bien se agrupan los

documentos dentro de estos tópicos.

A pesar de que los resultados de los procesos de *clustering* y clasificación permiten evaluar la calidad del modelo en términos de su capacidad para organizar documentos, esos resultados no fueron utilizadas para la selección del modelo óptimo, sino solo como una referencia adicional para analizar la calidad de los tópicos finales generados.

2.4. Módulo y *hardware* utilizados

Para el entrenamiento de los modelos de tópicos dinámicos DETM y CFDTM y su posterior evaluación, se utilizó el módulo de Python TopMost [1] (Wu, Pan y Luu 2023) que usa PyTorch (Paszke et al. 2019) como marco de trabajo de redes neuronales para modelos de tópicos neuronales. La elección de TopMost como marco de implementación se fundamenta en su facilidad de uso, su eficiencia computacional y su capacidad para trabajar con modelos de tópicos de última generación como DETM y CFDTM. Esta herramienta de código abierto proporciona una interfaz eficiente y flexible para la evaluación de los modelos y el preprocesamiento de los datos, lo que facilita el desarrollo y la comparación de diferentes enfoques de modelado de tópicos. La arquitectura de TopMost se muestra en la Figura 2.1.

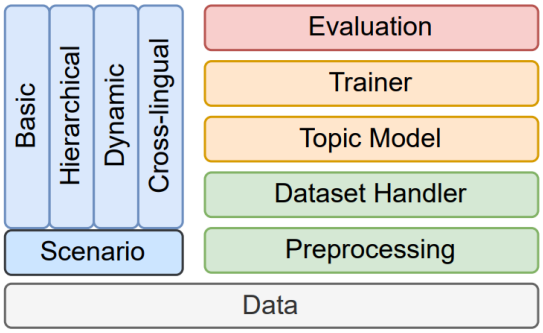

Figura 2.1: Arquitectura general del módulo TopMost. Cubre los escenarios más comunes de modelado de tópicos y desacopla la carga de datos, la construcción de modelos, el entrenamiento de modelos y las evaluaciones en los ciclos de vida del modelado de tópicos. (Wu, Pan y Luu 2023)

El entrenamiento de los modelos de tópicos dinámicos DETM y CFDTM se llevó a

[1] https://github.com/bobxwu/topmost

cabo utilizando la plataforma Kaggle, un entorno basado en la nube que proporciona recursos computacionales para el desarrollo de proyectos de aprendizaje automático. Kaggle ofrece acceso a diferentes tipos de *hardware*, incluyendo CPU y GPU, lo que permite acelerar el proceso de entrenamiento de modelos complejos.

Capítulo 3

Experimentos y Resultados

3.1. Modelo seleccionado

Para determinar el modelo que mejor se adapta al conjunto de datos, se realizaron múltiples experimentos con los modelos de tópicos dinámicos DETM y CFDTM sobre un corpus compuesto por artículos extraídos de la Revista de Ciencias Médicas de La Habana, Cuba. Se probaron diversas configuraciones de hiperparámetros y, tras cientos de experimentos, se encontraron las configuraciones óptimas que produjeron los mejores resultados. Dichas configuraciones se muestran en las tablas 3.1 y 3.2.

Tabla 3.1: Configuración óptima de hiperparámetros para el modelo DETM.

Hiperparámetro	Valor
Número de tópicos (`num_topics`)	20
Tamaño de ρ (`rho_size`)	400
Tamaño oculto de θ (`en_units`)	1200
Tamaño oculto de η (`eta_hidden_size`)	300
Capas en η (`eta_nlayers`)	4
Dropout en encoder (`enc_drop`)	0.2
Dropout en η (`eta_dropout`)	0.0
Delta (`delta`)	0.001
Activación en θ (`theta_act`)	SELU
Tasa de aprendizaje (`learning_rate`)	0.0005
Épocas (`epochs`)	400

Los resultados cuantitativos obtenidos en función de las métricas propuestas se presentan en la Tabla 3.3.

Tabla 3.2: Configuración óptima de hiperparámetros para el modelo CFDTM.

Hiperparámetro	Valor
Número de tópicos (`num_topics`)	20
Unidades en capa 1 (`en1_units`)	50
Dropout (`dropout`)	0.1
Temperatura beta (`beta_temp`)	1.5
Temperatura (`temperature`)	0.05
Peso negativo (`weight_neg`)	5e6
Peso positivo (`weight_pos`)	100
Peso UWE (`weight_UWE`)	1,000
Top-k negativo (`neg_topk`)	20
Tasa de aprendizaje (`learning_rate`)	0.001
Tamaño de *batch* (`batch_size`)	200
Épocas (`num_epoch`)	800

Tabla 3.3: Resultados cuantitativos de los modelos DETM y CFDTM.

Métrica	DETM	CFDTM
TC (`dynamic_TC`)	0.4587	0.5102
TD (`dynamic_TD`)	0.6520	0.9455
Pureza de Clustering (`Purity`)	0.2888	0.4210
Información Mutua Normalizada (`NMI`)	0.1481	0.3044
Exactitud en Clasificación (`acc`)	0.3252	0.4498
Macro-F1	0.1114	0.2725

Si bien la coherencia de los tópicos extraídos por ambos modelos es similar, se observa una diferencia notable en la diversidad de los tópicos generados. En particular, CFDTM muestra una diversidad significativamente mayor en comparación con DETM, lo que sugiere que este modelo captura de manera más efectiva los cambios drásticos en los tópicos latentes a lo largo del tiempo en los artículos de la Revista de Ciencias Médicas de La Habana. Estos resultados indican que CFDTM extrae tópicos con mayor calidad y representa mejor la evolución temática de la revista en diferentes períodos.

3.2. Tópicos extraídos

Para facilitar la interpretación de los resultados, por cada tópico se incluye un mapa de calor (*heatmap*), una representación gráfica en la que los valores se expresan mediante una escala de colores. En este caso, el *heatmap* permite visualizar de manera intuitiva cómo varía la probabilidad de cada palabra dentro del tópico a lo largo del tiempo, resaltando aquellas con mayor peso en cada año del período 2012-2024.

Además, cuando el contenido del tópico muestra una variación significativa a lo largo del tiempo, ya sea según la interpretación de los expertos o los resultados del modelo, se incluye una tabla que detalla cada tópico junto con sus palabras clave en los diferentes momentos en que fueron identificados.

Tópico#0

Las palabras clave identificadas entre 2012 y 2018 indican un enfoque en artículos científicos y estudios relacionados con la medicina ocupacional y la psicología médica, con términos como 'laboral', 'cuestionario' y 'nivel'.

Entre 2019 y 2021, el foco se amplió hacia el bienestar psicológico y el acceso abierto a la información científica, reflejado en palabras como 'burnout', 'anemia', 'acceso' y 'comerciales'.

A partir de 2021, los términos más relevantes se relacionan con la educación médica, especialmente en la evaluación del conocimiento de los estudiantes de medicina y la ética en la publicación científica, con énfasis en el acceso abierto y las licencias en medicina. En 2022, se mantiene este enfoque con términos como 'estudiantes', 'universitarios' y 'conocimiento'. Para 2023, la atención se ha desplazado hacia la salud digital y sus efectos, con palabras clave como 'fatiga', 'ruido', 'digitales' y 'videos'.

Tabla 3.4: Tópicos principales y palabras clave por año asociados al Tópico#0

Período	**tópico Principal**	**Palabras Clave**
2012-2018	Medicina ocupacional	laboral, cuestionario, nivel
2019-2021	Acceso abierto y bienestar psicológico	burnout, anemia, acceso, comerciales
2022	Educación médica	estudiantes, universitarios, conocimiento
2023	Salud digital	fatiga, ruido, digitales, videos

En síntesis, la evolución de este tópico transitó desde un enfoque en la medici-

na ocupacional (2012-2018) hacia el bienestar psicológico y el acceso abierto a la información (2019-2021). Posteriormente, centró su atención en la educación médica (2021-2022) y, más recientemente, en la salud digital y sus implicaciones (2023). (Tabla 3.4, Figura 3.1)

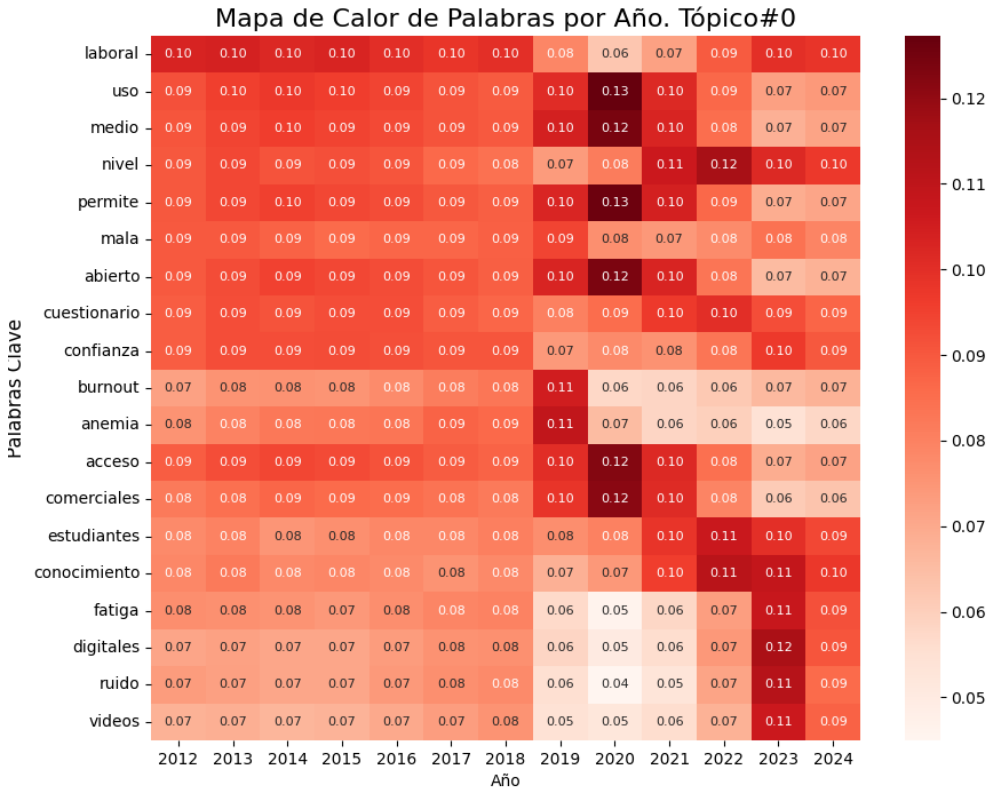

Figura 3.1: Mapa de calor de palabras por año (probabilidades) referentes al Tópico#0

Tópico#1 (Historia de la medicina en Cuba)

Las palabras reflejan que este tópico se mantuvo con variaciones insignificantes a lo largo del tiempo. Los términos de mayor relevancia, como 'historia', 'universidad', 'habana', 'medicina', 'cubana', 'cuba', 'cubanos' y 'doctor', indican que el tópico está relacionado con la historia de la medicina en Cuba. En algunos años, se destacan

personalidades significativas de la medicina, como el Dr. Alberto Juan Dorta Contreras, con los términos 'dorta' y 'contreras', así como referencias al Instituto Carlos J. Finlay mediante los términos 'instituto' y 'carlos'. (Figura 3.2).

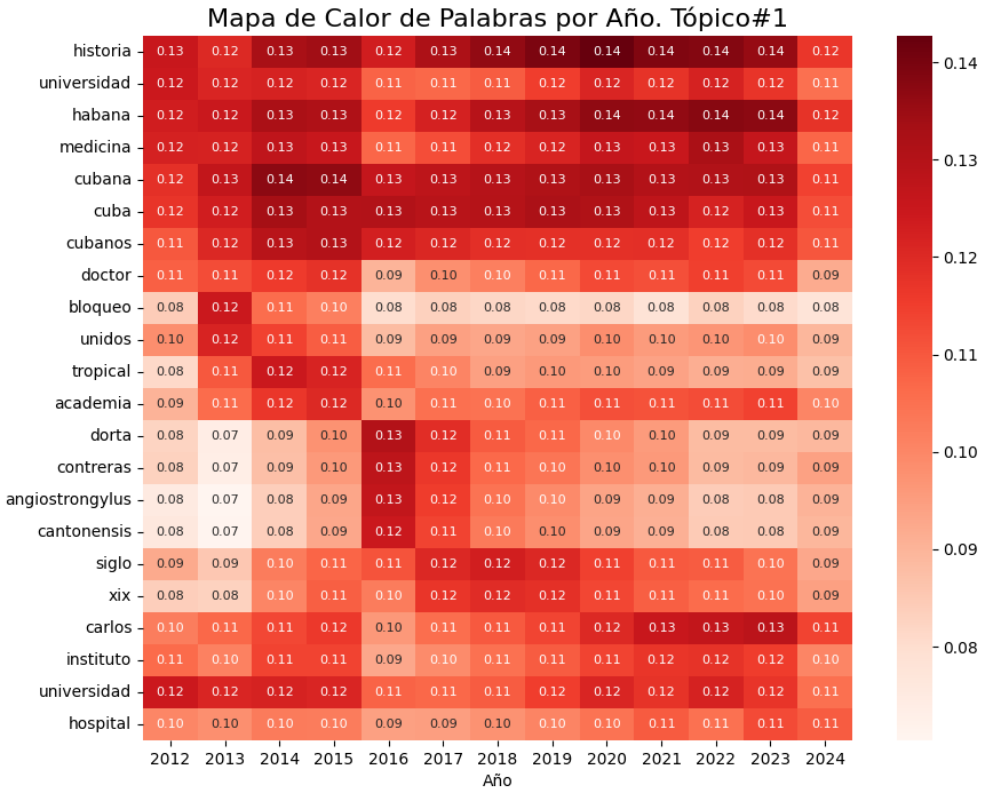

Figura 3.2: Mapa de calor de palabras por año (probabilidades) referentes al Tópico#1

Tópico#3 (Educación médica en Cuba)

El conjunto de términos relevantes en este tópico reflejan la evolución de la educación médica en Cuba, desde el diseño curricular y la planificación académica hasta la inclusión de estrategias de evaluación y el impacto de la salud pública. Los primeros años muestran un enfoque en la enseñanza de habilidades y planificación educativa,

mientras que en 2022 se observa un giro hacia la vacunación y estrategias sanitarias. Finalmente, en 2023 y 2024 se introduce la evaluación basada en competencias.

En el año 2012, se abordan aspectos de la planificación curricular y la enseñanza de habilidades en la educación médica con términos como 'contenidos', 'habilidades', 'estudiante', 'plan', 'alumnos', 'preguntas', 'carrera' e 'instrumento', lo que sugiere el desarrollo de programas académicos en medicina.

En el año 2013, los términos 'voz', 'lengua', 'sonidos', 'juegos', 'fuerza', 'english', 'sugerencias', 'clases' y 'palabras' indican un énfasis en la enseñanza del idioma y la comunicación médica.

En el período 2014-2019, términos como 'puntos', 'instrumento', 'ejercicio', 'estudiante', 'preguntas', 'estudiantes', 'aspectos', 'profesores', 'carrera' y 'encuesta' sugieren la implementación de evaluaciones educativas y encuestas docentes para medir el rendimiento académico, haciendo referencia también a las fases del aprendizaje en medicina con términos como 'cada', 'etapa', 'estudiante' y la inclusión de exámenes más rigurosos con los términos 'pruebas', 'entrenamiento' y 'dificultad', lo que indica un enfoque en la evaluación de habilidades clínicas

En el período 2020-2021, los términos 'recomendaciones', 'carrera' y 'plan' pueden estar relacionados con ajustes en los programas de estudio y recomendaciones en la formación médica.

En el período 2022, se observa un cambio significativo con la aparición de términos como 'vacuna', 'vacunas', 'paso', 'producto', 'estrategia' y 'etapa', lo que indica un interés en la salud pública y las estrategias de vacunación.

En el período 2023-2024, los términos 'asignatura', 'basadas' y 'examen' hacen referencia a la introducción de nuevas materias en el plan de estudio de la carrera de medicina y al seguimiento de metodologías de evaluación en la educación médica, posiblemente basadas en competencias o aprendizaje práctico.

En resumen, el análisis de estos términos sugiere una evolución en la educación médica en Cuba, desde la planificación curricular y el desarrollo de habilidades hasta la integración de estrategias de evaluación y el impacto de la salud pública en la enseñanza. (Tabla 3.5 y Figura 3.3)

Tópico#7 (VIH)

Las palabras clave por cada año en este tópico están relacionadas con diferentes aspectos del manejo del VIH en Cuba, con un enfoque particular en su impacto en la salud mental del paciente y el papel de la familia.

En el año 2012, se habla sobre la atención familiar y social hacia los pacientes con VIH. Las palabras como 'sida', 'vih', 'sexual', 'familiar', 'familia', 'vida', 'padres',

Tabla 3.5: Tópicos principales y palabras clave por año asociados al Tópico#3

Período	**Tópico Principal**	**Palabras Clave**
2012	Plan de estudio y habilidades	contenidos, habilidades, estudiante, plan, alumnos, preguntas, carrera, instrumento
2013	Enseñanza del idioma y comunicación médica	voz, lengua, sonidos, juegos, fuerza, english, sugerencias, clases, palabras
2014-2019	Evaluación educativa y medición del rendimiento académico	puntos, instrumento, ejercicio, estudiante, preguntas, aspectos, profesores, carrera, encuesta, cada, etapa, pruebas, entrenamiento, dificultad
2020-2021	Ajustes en programas de estudio y formación médica	recomendaciones, carrera, plan
2022	Estrategias de vacunación y salud pública	vacuna, vacunas, paso, producto, estrategia, etapa
2023-2024	Nuevas asignaturas y evaluación basada en competencias	asignatura, basadas, examen

'mujeres', 'ancianos' sugieren estudios sobre el impacto del VIH en la familia, el entorno social y los cuidados necesarios para pacientes con enfermedades crónicas.

En el período 2013-2017, se habla sobre la psicología y el envejecimiento en pacientes con VIH, con términos como 'envejecimiento', 'suicidio', 'relaciones', 'adolescentes', 'personas', 'sexual', 'consumo', que indican estudios sobre las reacciones psicológicas de los pacientes con VIH ante el deterioro de su estado de salud, el envejecimiento prematuro y el riesgo de suicidio. También se observan referencias a estilos de vida y consumo de sustancias.

En el año 2018, se habla sobre la estigmatización y la salud mental en pacientes con VIH, con términos como 'mental', 'estigma', 'conductas', que sugieren investigaciones sobre la discriminación hacia las personas con VIH y el impacto en su bienestar psicológico.

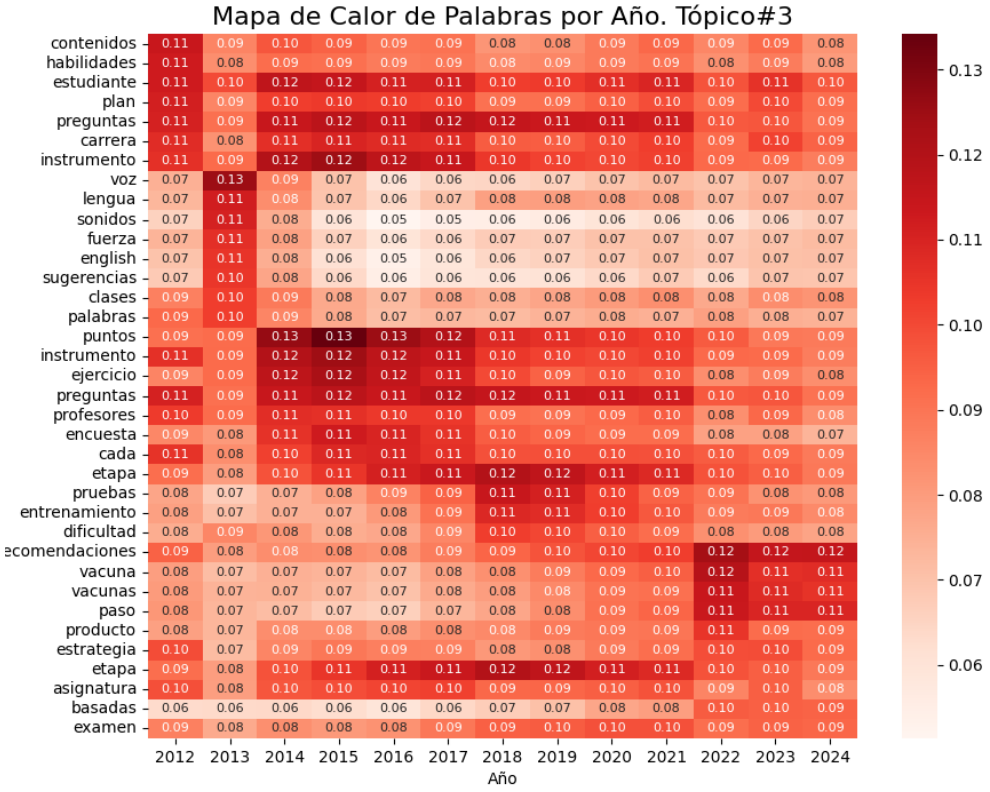

Figura 3.3: Mapa de calor de palabras por año (probabilidades) referentes al Tópico#3

Para concluir, en el período 2019-2024 se habla sobre aspectos psicológicos y salud mental en pacientes con VIH. Durante este período, las palabras como 'mental', 'cuidadores', 'suicidio', 'trastorno', 'bipolar', 'intento', 'alcohol', 'adolescentes', 'demencia', 'estilos', 'conducta', 'adultos' muestran un enfoque en los cambios psicológicos en personas con VIH, incluyendo trastornos mentales, intentos de suicidio y el rol de los cuidadores en el manejo de la enfermedad.

En resumen, a partir de 2012, los estudios evolucionaron desde el impacto del VIH en la familia hasta los aspectos psicológicos de los pacientes. Se observa un interés creciente en la salud mental, el suicidio y el estigma social. En los últimos años (2019-2024), los estudios se centraron en la progresión de problemas psicológicos

en pacientes con VIH, con términos relacionados con trastornos mentales, el consumo de sustancias y el rol de los cuidadores. (Tabla 3.6, Figura 3.4)

Tabla 3.6: Tópicos principales y palabras clave por año asociados al Tópico#7

Período	**Tópico Principal**	**Palabras Clave**
2012	Atención familiar y social los pacientes	sida, vih, familiar, sexual, familia, vida, padres, mujeres, ancianos
2013-2017	Psicología y envejecimiento en pacientes con VIH	sexual, suicidio, relaciones, adolescentes, personas, envejecimiento, consumo
2018	Estigmatización y salud mental en pacientes con VIH	mental, estigma, conductas
2019-2024	Aspectos psicológicos y salud mental en pacientes con VIH	mental, cuidadores, suicidio, trastorno, bipolar, intento, alcohol, adolescentes, demencia, estilos, conducta, adultos

Tópico#8 (Epidemiología)

Las palabras clave asociadas a este tópico reflejan la evolución de las investigaciones médicas, enfocándose en la vigilancia epidemiológica, el control de enfermedades infecciosas y las respuestas sanitarias a crisis emergentes. A partir de 2021, la pandemia de COVID-19 tomó un papel central, con investigaciones sobre su impacto, resistencia antimicrobiana y estrategias de vacunación.

En el período 2012-2013, los estudios se centraron en la epidemiología y la vigilancia epidemiológica en la salud pública, abordando las fuentes de transmisión y el control de enfermedades, con términos como 'vigilancia', 'datos', 'casos', 'provincia', 'centros', 'fallecidos', 'red', 'disponible', 'muestras' y 'fuente'.

En el período 2014-2015, las investigaciones se enfocaron en las intoxicaciones, las consultas médicas, las enfermedades agudas y las emergencias, con términos como 'integrado', 'intoxicaciones', 'ocurrencia', 'consultas' y 'agudos'.

En el período 2016-2020, los términos con mayor relevancia sugieren que se realizaron investigaciones sobre mortalidad, análisis post-mortem, centros de salud, fuentes

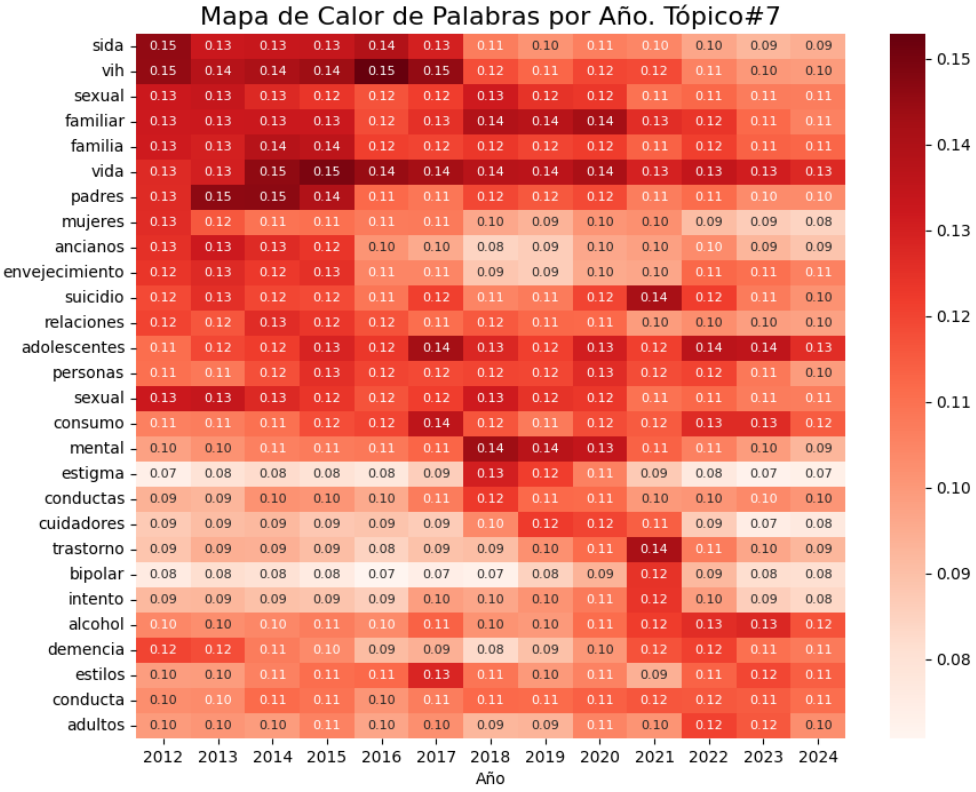

Figura 3.4: Mapa de calor de palabras por año (probabilidades) referentes al Tópico#7

de contagio y estudios sobre el dengue y los brotes epidémicos, con términos como 'muestras', 'fallecidos', 'centros', 'fuente', 'alarma', 'dengue', 'Lambayeque'.

Finalmente, en el período de 2021-2024, los estudios se centraron en la pandemia de COVID-19, las enfermedades cardiovasculares, la resistencia antimicrobiana y los modelos epidemiológicos, con términos como 'pandemia', 'resistencia', 'impacto', 'cubanas', 'vigilancia', 'covid', 'competitividad', 'gobiernos', 'coronavirus', 'mortalidad', 'cov', 'sars', 'disponible', 'modelos', 'internet' y 'dosis'.

En resumen, este tópico evolucionó desde un enfoque en la vigilancia epidemiológica hasta el impacto y control de la pandemia de COVID-19, con un creciente uso de herramientas digitales y modelos predictivos en la salud pública. (Tabla 3.7,

Figura 3.5)

Tabla 3.7: Tópicos principales y palabras clave por año asociados al Tópico#8

Período	**Tópico Principal**	**Palabras Clave**
2012-2013	Epidemiología y la vigilancia epidemiológica: Fuentes de transmisión y el control de enfermedades	vigilancia, datos, casos, provincia, centros, fallecidos, red, disponible, muestras, fuente
2014-2015	Intoxicaciones, Consultas médicas, Enfermedades agudas	intoxicaciones, ocurrencia, consultas, agudos
2016-2020	Mortalidad, Análisis postmortem, Centros de salud, Fuentes de contagio	muestras, fallecidos, centros, fuente, alarma, dengue, Lambayeque
2021-2024	COVID-19, Enfermedades cardiovasculares, Resistencia antimicrobiana , Modelos epidemiológicos	pandemia, resistencia, impacto, cubanas, vigilancia, covid, competitividad, gobiernos, coronavirus, mortalidad, cov, sars, modelos, internet, dosis

Tópico#9 (Enfermedades abdominales e intestinales)

Este tópico no presenta variaciones significativas en el período 2012-2024. Los estudios se centraron principalmente en enfermedades abdominales e intestinales, empezando con el diagnóstico y manejo de afecciones frecuentes, pasando por la identificación de quistes y su manejo mediante biopsias. A partir de 2015, el enfoque cambió hacia estudios sobre el carcinoma (posiblemente cáncer) abdominal o intestinal. En 2016, se observó un enfoque en el diagnóstico genético o molecular de enfermedades raras. La tendencia en los últimos años siguió siendo en torno a las complicaciones de las enfermedades abdominales y el diagnóstico mediante biopsias, con un énfasis continuo en la fiebre como síntoma en 2024. Los términos más relevantes se muestran en la Figura 3.6.

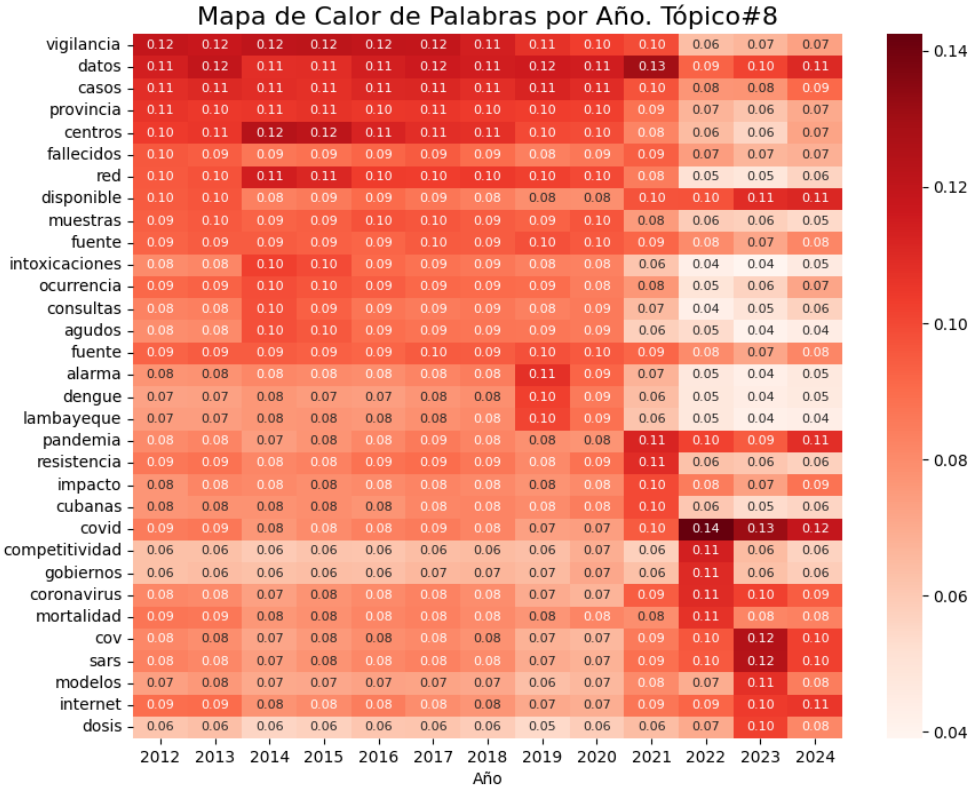

Figura 3.5: Mapa de calor de palabras por año (probabilidades) referentes al Tópico#8

Tópico#10

Las palabras clave asociadas a este tópico sugieren que en 2012, los estudios se centraron en las complicaciones relacionadas con Helicobacter pylori y linfomas, utilizando términos como 'pylori', 'seguimiento', 'helicobacter', 'reportado', 'temporal', 'linfoma' y 'manifestaciones'.

Entre 2013-2014, los estudios se enfocaron en el virus Varicela Zóster y las complicaciones asociadas, incluidas las manifestaciones nerviosas, con términos como 'temporal', 'varicela', 'zoster', 'serial', 'gigantes', 'fetal', 'nervio' y 'mano'.

En el período 2015, se realizaron estudios sobre el síndrome de ojo seco y las complicaciones del nervio mediano, con términos como 'ojo', 'síndrome', 'visual',

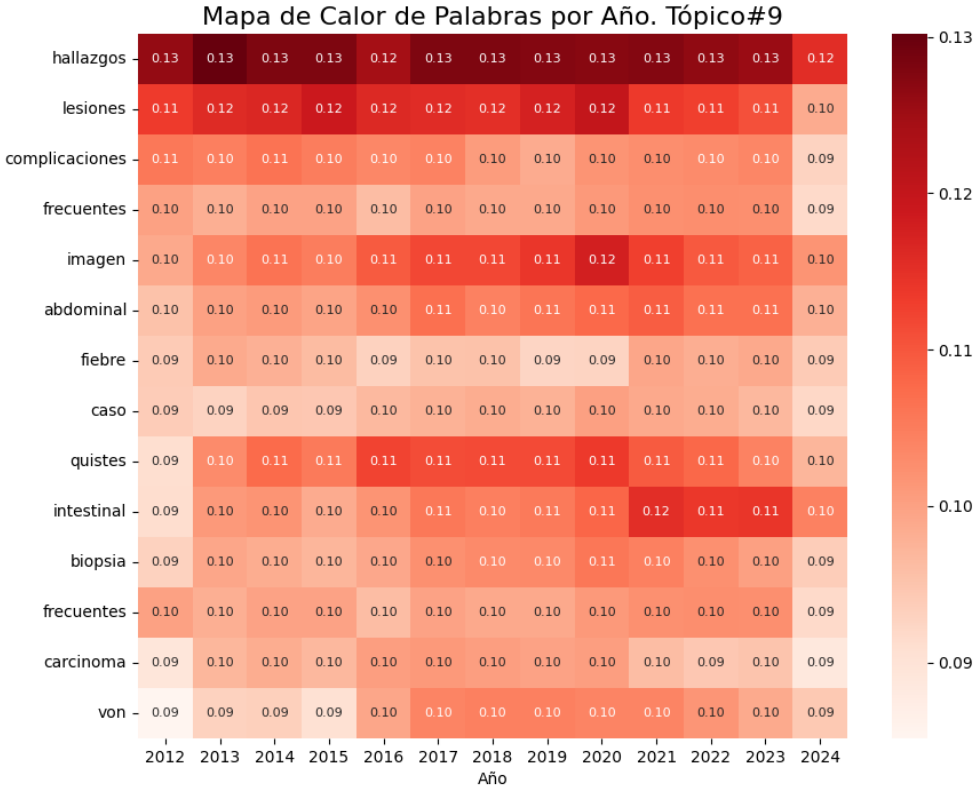

Figura 3.6: Mapa de calor de palabras por año (probabilidades) referentes al Tópico#9

'seco', 'Johnson', 'mediano', 'superficie', 'temporal', 'nervio' y 'seguimiento'.

Entre 2016-2018, el enfoque cambió al síndrome de Stevens-Johnson, con complicaciones oculares bilaterales, seguido por estudios sobre el seguimiento de estas enfermedades a lo largo del tiempo. Este tópico se detectó mediante los términos 'seco', 'Stevens', 'bilateral', 'mediano' y 'corneal'.

En el último período, 2019-2024, el enfoque fue más general, centrado en los cuadros clínicos de diversas patologías. En 2021, se observó la aparición de nuevas entidades clínicas y finalmente, en 2024, el interés se focalizó en las consultas para el diagnóstico y manejo de enfermedades relacionadas con los ojos o el sistema nervioso. (Tabla 3.8, Figura 3.7)

Tabla 3.8: Tópicos principales y palabras clave por año asociados al Tópico#10

Período	**Tópico Principal**	**Palabras Clave**
2012	Helicobacter pylori y linfomas	pylori, seguimiento, helicobacter, reportado, temporal, linfoma, manifestaciones
2013-2014	Virus varicela zóster y complicaciones asociadas	temporal, varicela, zoster, serial, gigantes, fetal, nervio, mano
2015	Síndrome de ojo seco y complicaciones del nervio mediano	ojo, síndrome, visual, seco, Johnson, mediano, superficie, temporal, nervio, seguimiento
2016-2024	Cuadros clínicos y consultas para diagnóstico	cuadro, seguimiento, aparece, entidad, temporal, consulta

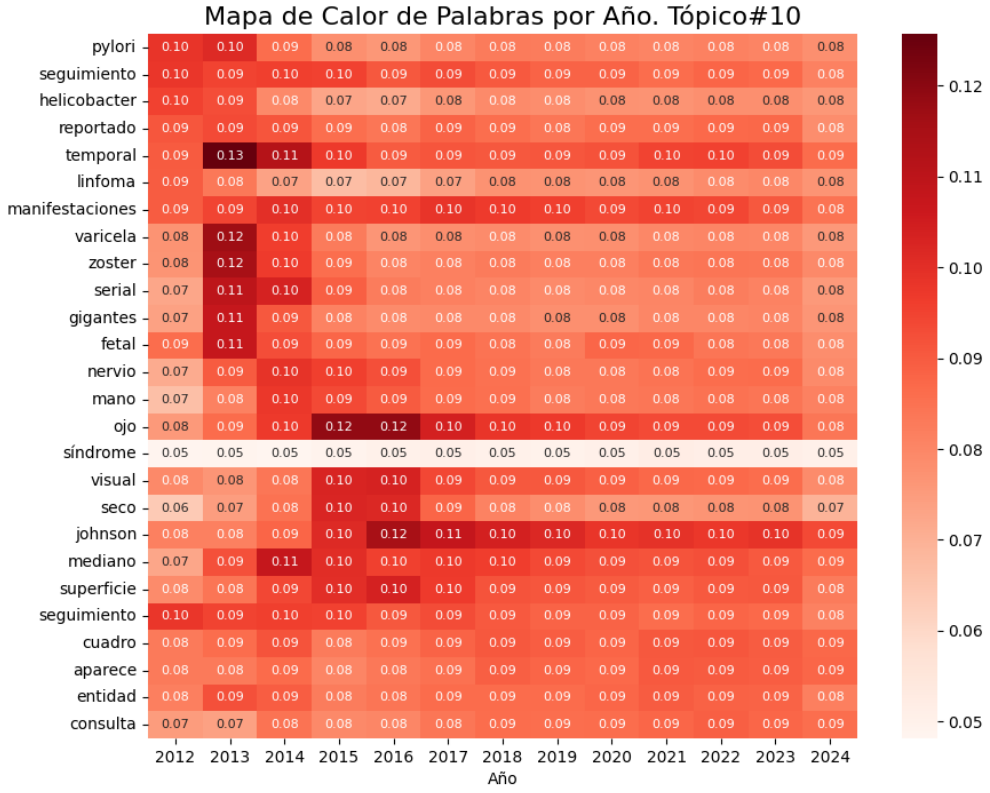

Figura 3.7: Mapa de calor de palabras por año (probabilidades) referentes al Tópico#10

Conclusiones

El presente estudio analizó y comparó el desempeño de los modelos de tópicos dinámicos DETM y CFDTM en el análisis de los artículos publicados en la Revista de Ciencias Médicas de La Habana, Cuba. A lo largo del trabajo, se realizaron experimentos extensivos con distintas configuraciones de hiperparámetros con el objetivo de determinar cuál de los dos modelos lograba representar de manera más efectiva la evolución de los tópicos en el tiempo.

Los resultados obtenidos indican que, si bien ambos modelos presentaron un rendimiento similar en términos de coherencia de tópicos, CFDTM demostró una ventaja significativa en términos de diversidad temática. La capacidad de este modelo para extraer tópicos más diferenciados y menos redundantes sugiere que representa mejor los cambios temáticos a lo largo del tiempo. Asimismo, los experimentos de clasificación y agrupamiento de documentos evidenciaron que CFDTM permite una mejor organización de los textos en función de sus temáticas, lo que facilita la identificación de tendencias emergentes en la investigación médica.

Otro aspecto relevante del estudio fue la búsqueda de configuraciones óptimas de hiperparámetros, lo que permitió mejorar notablemente el rendimiento de ambos modelos. La selección adecuada de los hiperparámetros resultó crucial para maximizar la calidad de los tópicos extraídos.

En conclusión, los hallazgos de esta investigación sugieren que el modelo CFDTM es más adecuado que DETM para analizar la evolución de los tópicos en revistas científicas. Su capacidad para generar tópicos más diversos y estructurados lo convierte en una herramienta valiosa para el estudio de la dinámica temática en publicaciones científicas especializadas.

Recomendaciones

A partir de los resultados obtenidos en este estudio, se proponen varias líneas de trabajo futuro que podrían contribuir al desarrollo y la optimización de los modelos de tópicos dinámicos.

En primer lugar, sería interesante explorar la fusión del modelo CFDTM con el modelo TopicGPT (Pham et al. 2023) en lugar de LDA para la captura de tópicos en un momento específico en el tiempo. Dado que TopicGPT ha demostrado capacidades avanzadas en la generación de representaciones semánticas, su integración con CFDTM podría mejorar la calidad y precisión de los tópicos extraídos, proporcionando una mejor interpretación de la evolución temática en conjuntos de datos de naturaleza dinámica.

Asimismo, se recomienda la creación de una herramienta que, una vez entrenados los modelos de tópicos dinámicos, sea capaz de reconstruir visualmente la evolución de los tópicos a lo largo del tiempo. Este tipo de herramienta facilitaría la interpretación y el análisis de tendencias en investigaciones científicas, ofreciendo a los especialistas una forma más intuitiva de comprender la dinámica de los tópicos abordados en la literatura académica.

Finalmente, se sugiere ampliar la evaluación de los modelos incluyendo métricas adicionales que midan el impacto práctico de los tópicos extraídos en tareas específicas, como la recuperación de información y la recomendación de artículos científicos. La combinación de enfoques cualitativos y cuantitativos en la evaluación contribuiría a una mejor selección del modelo más adecuado según el propósito del análisis.

Bibliografía

Adji B. Dieng, D. M. B., Francisco J. R. Ruiz. (2019). The Dynamic Embedded Topic Model. https://arxiv.org/abs/1907.05545 (vid. págs. 2, 6, 8).

Song, D. J. C. Z. Y. (2023). *Probabilistic Topic Models. Foundation and Application*.

Wu, X., Dong, X., Pan, L., Nguyen, T., & Luu, A. T. (2024). Modeling Dynamic Topics in Chain-Free Fashion by Evolution-Tracking Contrastive Learning and Unassociated Word Exclusion. https://arxiv.org/abs/2405.17957 (vid. págs. 2, 8).

Wu, X., Nguyen, T., & Luu, A. T. (2024). A Survey on Neural Topic Models: Methods, Applications, and Challenges. *ArXiv*, *abs/2401.15351*. https://api.semanticscholar.org/CorpusID:267297321 (vid. pág. 8).

Wu, X., Pan, F., & Luu, A. T. (2023). Towards the TopMost: A Topic Modeling System Toolkit. *ArXiv*, *abs/2309.06908*. https://api.semanticscholar.org/CorpusID:261705592 (vid. págs. 10, 17).

Referencias

Adji B. Dieng, D. M. B., Francisco J. R. Ruiz. (2019). The Dynamic Embedded Topic Model. https://arxiv.org/abs/1907.05545 (vid. págs. 2, 6, 8).

Arthur P. Dempster, D. B. R., Nan M. Laird. (1977). Maximum likelihood from incomplete data via the EM - algorithm plus discussions on the paper. https://api.semanticscholar.org/CorpusID:4193919 (vid. pág. 4).

Baevski, A., Hsu, W.-N., Xu, Q., Babu, A., Gu, J., & Auli, M. (2022). data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language. *ArXiv*, *abs/2202.03555*. https://api.semanticscholar.org/CorpusID:246652264 (vid. pág. 7).

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, *112*(518), 859-877. https://doi.org/10.1080/01621459.2017.1285773 (vid. pág. 6).

Blei, D. M., Ng, A., & Jordan, M. I. (2003). Latent Dirichlet Allocation. https://api.semanticscholar.org/CorpusID:3177797 (vid. págs. 4, 6).

Blei & Lafferty. (2006). Dynamic topic models. https://doi.org/10.1145/1143844.1143859 (vid. págs. 6, 8).

Cvejoski, K., Sánchez, R. J., & Ojeda, C. (2023). Neural Dynamic Focused Topic Model. https://arxiv.org/abs/2301.10988 (vid. pág. 7).

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Fumas, G. W., & Beck, L. L. (1988). Improving information retrieval using latent semantic indexing. https://api.semanticscholar.org/CorpusID:59739393 (vid. pág. 4).

Dieng, A. B., Ruiz, F. J. R., & Blei, D. M. (2019). Topic Modeling in Embedding Spaces. *Transactions of the Association for Computational Linguistics*, *8*, 439-453. https://api.semanticscholar.org/CorpusID:195886143 (vid. pág. 6).

Gershman, S. J., & Goodman, N. D. (2014). Amortized Inference in Probabilistic Reasoning. *Cognitive Science*, *36*. https://api.semanticscholar.org/CorpusID:924780 (vid. pág. 6).

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*, 1735-1780. https://api.semanticscholar.org/CorpusID:1915014 (vid. pág. 6).

Hoffman, M., Blei, D. M., Wang, C., & Paisley, J. (2013). Stochastic Variational Inference. https://arxiv.org/abs/1206.7051 (vid. pág. 6).

Hofmann, T. (1999). Probabilistic Latent Semantic Analysis. *Conference on Uncertainty in Artificial Intelligence*, 289-296. https://api.semanticscholar.org/CorpusID:653762 (vid. pág. 4).

Jordan, M. I., Ghahramani, Z., Jaakkola, T., & Saul, L. K. (1999). An Introduction to Variational Methods for Graphical Models. *Machine Learning*, *37*, 183-233. https://api.semanticscholar.org/CorpusID:2073260 (vid. pág. 6).

Li, W., & McCallum, A. (2006). Pachinko allocation: DAG-structured mixture models of topic correlations. *Proceedings of the 23rd international conference on Machine learning*, 577-584. https://api.semanticscholar.org/CorpusID:13160178 (vid. pág. 5).

Lin, Z., Feng, M., dos Santos, C. N., Yu, M., Xiang, B., Zhou, B., & Bengio, Y. (2017). A Structured Self-attentive Sentence Embedding. https://arxiv.org/abs/1703.03130 (vid. pág. 7).

Miyamoto, N., Isonuma, M., Takase, S., Mori, J., & Sakata, I. (2023, julio). Dynamic Structured Neural Topic Model with Self-Attention Mechanism. En A. Rogers, J. Boyd-Graber & N. Okazaki (Eds.), *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 5916-5930). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.366 (vid. págs. 7, 8).

Murphy, K. P. (2012). Machine learning - a probabilistic perspective. *Adaptive computation and machine learning series*. https://api.semanticscholar.org/CorpusID:17793133 (vid. pág. 5).

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., ... Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. *ArXiv*, *abs/1912.01703*. https://api.semanticscholar.org/CorpusID:202786778 (vid. pág. 17).

Pham, C. M., Hoyle, A. M., Sun, S., & Iyyer, M. (2023). TopicGPT: A Prompt-based Topic Modeling Framework. *ArXiv*, *abs/2311.01449*. https://api.semanticscholar.org/CorpusID:272144718 (vid. pág. 35).

Rahimi, H., Naacke, H., Constantin, C., & Amann, B. (2023). ANTM: An Aligned Neural Topic Model for Exploring Evolving Topics. https://arxiv.org/abs/2302.01501 (vid. pág. 7).

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. https://api.semanticscholar.org/CorpusID:7743332 (vid. pág. 16).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). Attention Is All You Need. https://arxiv.org/abs/1706.03762 (vid. pág. 7).

Wang, X., & McCallum, A. (2006). Topics over time: a non-Markov continuous-time model of topical trends. *Knowledge Discovery and Data Mining*, 424-433. https://api.semanticscholar.org/CorpusID:207160148 (vid. pág. 5).

Wu, X., Dong, X., Pan, L., Nguyen, T., & Luu, A. T. (2024). Modeling Dynamic Topics in Chain-Free Fashion by Evolution-Tracking Contrastive Learning and Unassociated Word Exclusion. https://arxiv.org/abs/2405.17957 (vid. págs. 2, 8).

Zhang, D. C., & Lauw, H. (2022, 17–23 Jul). Dynamic Topic Models for Temporal Document Networks. En K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu & S. Sabato (Eds.), *Proceedings of the 39th International Conference on Machine Learning* (pp. 26281-26292, Vol. 162). PMLR. https://proceedings.mlr.press/v162/zhang22n.html (vid. pág. 6).

Anexos

Lista de tópicos más relevantes extraídos (2012 - 2024)

Tópico#0

2012: laboral uso medio nivel permite mala abierto cualquier acceso cuestionario
2013: laboral uso medio permite nivel abierto cualquier acceso cuestionario confianza
2014: laboral uso medio permite abierto issn acceso cualquier media confianza
2015: laboral uso issn medio media permite abierto recibido acceso confianza
2016: laboral uso media medio permite abierto cuestionario confianza issn acceso
2017: laboral medio uso media confianza abierto permite university acceso students
2018: laboral university confianza medio universitarios uso abierto permite
2019: anemia ref burnout medio abierto permite cualquier uso apego licenses
2020: uso permite medio cualquier abierto acceso comerciales commons comercial licenses
2021: nivel permite medio abierto cualquier uso acceso licenses commons comerciales
2022: nivel conocimiento regular estudiantes universitarios conocimientos cuestionario education
2023: universitarios digitales desacuerdo digital ruido entornos acuerdo conocimiento fatiga videos
2024: deportivas sports universitarios acuerdo chimborazo deporte laboral falta contingencia

Tópico#1

2012: historia universidad habana medicina cubana cuba presidente cubanos doctor
2013: cubana habana bloqueo cuba medicina unidos universidad cubanos historia nace
2014: cubana cuba historia habana cubanos medicina tropical universidad sociedad academia
2015: cubana historia habana cubanos cuba medicina tropical universidad academia doctor
2016: cuba dorta contreras angiostrongylus cubana cantonensis historia citaciones cubanos habana

2017: historia cuba cubana habana cubanos siglo dorta xix contreras angiostrongylus
2018: historia cuba cubana habana siglo experimental xix medicina cubanos cuban
2019: historia habana cuba cubana medicina siglo cubanos xix experimental universidad
2020: historia habana cubana cuba medicina universidad carlos cubanos siglo instituto
2021: historia habana cubana cuba medicina carlos cubanos universidad instituto cuban
2022: historia habana medicina cubana carlos cuba universidad instituto cubanos doctor
2023: habana historia cubana carlos medicina cuba francisco cubanos universidad instituto
2024: habana historia cubana carlos cuba cubanos hospital francisco medicina universidad

Tópico#3

2012: casi nunca contenidos habilidades estudiante plan alumnos preguntas carrera instrumento
2013: voz lengua sonidos juegos fuerza teaching english sugerencias clases palabras
2014: puntos instrumento ejercicio estudiante preguntas estudiantes aspectos profesores carrera encuesta
2015: puntos instrumento ejercicio preguntas aspectos estudiante estudiantes encuesta educativa instrumentos
2016: puntos instrumento ejercicio preguntas aspectos etapa estudiante estudiantes encuesta cada
2017: puntos preguntas instrumento etapa estudiante ejercicio estudiantes aspectos carrera cada
2018: etapa preguntas pruebas entrenamiento puntos instrumento estudiante dificultad estudiantes
2019: etapa preguntas pruebas entrenamiento puntos estudiante instrumento dificultad aspectos pregunta
2020: preguntas etapa estudiante puntos entrenamiento pruebas recomendaciones aspectos instrumento carrera
2021: preguntas estudiante etapa puntos carrera recomendaciones instrumento aspectos plan pregunta
2022: recomendaciones vacuna vacunas paso producto estrategia etapa estudiante preguntas puntos
2023: recomendaciones paso vacunas vacuna estudiante carrera plan estrategia preguntas asignatura
2024: recomendaciones paso vacuna vacunas estudiante carrera basadas examen estrategia plan

Tópico#7

2012: sida vih sexual familiar familia sexuales vida padres mujeres ancianos
2013: padres vih sexual sida familiar ancianos vida familia envejecimiento suicidio
2014: padres vida vih familia sida familiar sexual ancianos relaciones envejecimiento
2015: vida vih padres familia sida familiar adolescentes personas envejecimiento sexual
2016: vih vida sida life adolescentes familia personas consumo familiar relaciones
2017: vih adolescentes vida consumo sida estilos familiar sexual suicidio familia
2018: mental familiar vida sexual estigma adolescentes people familia conductas familias
2019: familiar mental vida sexual cuidadores estigma adolescentes cuidador familia personas
2020: familiar vida mental adolescentes personas sexual familia cuidador cuidadores suicidio

2021: suicidio trastorno vida suicida intentos familiar bipolar intento suicide alcohol
2022: adolescentes vida adolescente consumo alcohol familiar mayores suicidio familia demencia
2023: adolescentes adolescente vida alcohol consumo estilos mayores adultos conducta adolescencia
2024: adolescente vida adolescentes alcohol adolescencia consumo estilos familia conducta estilo

Tópico#8

2012: vigilancia datos casos provincia centros fallecidos red disponible citado muestras
2013: datos vigilancia casos centros provincia red disponible citado muestras fuente
2014: centros vigilancia red casos datos provincia integrado intoxicaciones ocurrencia consultas
2015: centros vigilancia red datos casos provincia integrado intoxicaciones ocurrencia agudos
2016: vigilancia datos centros casos provincia red muestras ocurrencia integrado fallecidos
2017: vigilancia datos casos centros provincia red muestras fuente fallecidos integrado
2018: datos vigilancia casos centros red provincia fuente muestras integrado ocurrencia
2019: datos casos alarma vigilancia provincia dengue red lambayeque centros fuente
2020: datos casos vigilancia provincia red centros muestras laboratorio fuente alarma
2021: datos data pandemia resistencia impacto cubanas vigilancia citado
2022: covid competitividad mortality gobiernos coronavirus ica mortalidad
2023: covid cov sars citado disponible modelos internet coronavirus dosis
2024: citado covid disponible datos pandemia internet cov sars

Tópico#9

2012: hallazgos lesiones complicaciones frecuentes imagen abdominal fiebre caso
2013: hallazgos lesiones complicaciones imagen quistes intestinal abdominal biopsia fiebre
2014: hallazgos lesiones quistes complicaciones imagen intestinal abdominal frecuentes
2015: hallazgos lesiones quistes complicaciones imagen frecuentes abdominal intestinal
2016: hallazgos lesiones quistes imagen complicaciones abdominal intestinal carcinoma
2017: hallazgos lesiones imagen quistes abdominal intestinal complicaciones von
2018: hallazgos lesiones imagen quistes intestinal von abdominal biopsia
2019: hallazgos lesiones imagen quistes abdominal intestinal von biopsia
2020: hallazgos lesiones imagen quistes intestinal abdominal biopsia von
2021: hallazgos intestinal lesiones imagen quistes abdominal intestino biopsia von
2022: hallazgos intestinal lesiones imagen quistes abdominal intestino complicaciones
2023: hallazgos intestinal lesiones imagen intestino abdominal quistes complicaciones frecuentes
2024: hallazgos intestinal intestino imagen lesiones abdominal quistes von fiebre

Tópico#10

2012: pylori seguimiento estadio helicobacter reportado temporal linfoma manifestaciones pediatr
2013: temporal varicela zoster february serial gigantes fetal giant
2014: temporal mediano serial manifestaciones nervio mano ojo zoster
2015: ojo visual seco johnson mediano superficie temporal nervio seguimiento
2016: ojo johnson superficie visual seco stevens bilateral mediano corneal
2017: johnson ojo manifestaciones mediano superficie visual seguimiento stevens coincidiendo
2018: johnson ojo manifestaciones mediano coincidiendo superficie cuadro visual temporal
2019: johnson ojo manifestaciones mediano coincidiendo superficie temporal visual cuadro
2020: johnson ojo manifestaciones mediano temporal coincidiendo superficie visual seguimiento
2021: johnson manifestaciones temporal ojo coincidiendo mediano cuadro aparece entidad
2022: johnson temporal manifestaciones ojo coincidiendo mediano cuadro superficie encontramos
2023: johnson coincidiendo temporal ojo encontramos manifestaciones superficie mediano cuadro
2024: johnson coincidiendo aparece cuadro encontramos mediano temporal consulta manifestaciones