

Estado del Arte (Cap 2)

Los Modelos Tópicos Neuronales Dinámicos, DNTM por sus siglas en inglés, han evolucionado principalmente en la última década. Las variantes propuestas han alcanzado 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 Embedding Topic Model (D-ETM)* (Adji B. Dieng 2019) que sigue un enfoque diferente representando a palabras y tópicos como *embeddings* ². Este modelo tiene como bases a los modelos LDA (Blei, Ng, and Jordan 2003), *Dynamic LDA (D-LDA)* (Blei and Lafferty 2006) y *Embeddings Topic Model (ETM)* (Dieng, Ruiz, and Blei 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 mismas 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 de 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 aproximados mediante el mecanismo de inferencia variacional (Jordan et al. 1999; Blei, Kucukelbir, and McAuliffe 2017). Para escalar el algoritmo a grandes conjuntos de datos, fueron utilizados el submuestreo de datos (Hoffman et al. 2013) y la amortización (Gershman and Goodman 2014) con un parámetro de aproximación variacional estructurado parametrizado por una red neuronal de memoria a corto y largo plazo (LSTM) (Hochreiter and Schmidhuber 1997).

¹En 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.

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

En 2022 fue publicado el artículo Zhang and Lauw 2022 donde se describen dos modelos 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 cómo los documentos creados en diferentes momentos pueden formar vínculos. Dado que una red de documentos no emerge en su totalidad de forma imprevista, es posible afirmar que teniendo una red inicial 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 ventana deslizante superpuesta para la agrupación temporal de documentos. Este algoritmo, llamado *Aligned Clustering*, consiste en segmentar archivos en segmentos superpuestos, realizar agrupación secuencial basada 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, 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), 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 cómo 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 conjuntamente, 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 tópico puede ser algo ambiguo. Un ejemplo de esto es el modelo *Neural Dynamic Focused Topic Model* (NDF-TM) (Cvejowski, Sánchez, and Ojeda 2023), un modelo que en lugar de modelar la evolución de los tópicos modela las actividades de estos a lo largo del tiempo. Cabe destacar que las actividades 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 del modelado tópico dinámico (Wu, Nguyen, and Luu 2024). En este enfoque se busca desacoplar la probabilidad de que un tópico esté activo de su proporción mediante la introducción de secuencias de variables aleatorias de Bernoulli, que seleccionan los tópicos activos para un documento dado en un instante particular de tiempo. Para lograr el desacoplamiento sigue una lógica similar a la de los priors no paramétricos, como el proceso de buffet indio sobre matrices binarias infinitas, tanto en entornos estáticos (Williamson et al. 2010) como dinámicos (Perrone et al. 2016), pero aprovecha el truco de reparametrización para realizar inferencia variacional neural (Kingma and Welling 2022). El resultado es un modelo escalable que permite que el número instantáneo de tópicos activos por documento fluctúe y que desacople explícitamente la proporción de tópicos de su actividad, 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 repetidos³ y ubicar tópicos en espacios de tiempo incorrectos. 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 (Wu et al. 2024). Para resolver los problemas anteriores surge el modelo *Chain-Free Dynamic Topic Model* (CFDTM) (Wu 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, donde para rastrear la evolución del tema con diferentes intensidades construye relaciones positivas y más importante aún, construye relaciones negativas que 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 relacionadas de los tópicos en cada segmento de tiempo e identifica cuáles de ellas no pertenecen a este 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 and Lafferty 2006; Adji B. Dieng 2019; Miyamoto et al. 2023), 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 evolucio-

³Los tópicos dentro de un segmento de tiempo se considerarán repetitivos al presentar una semántica similar

nan 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 respecto al resto de los modelos existentes.

En conclusión, los Modelos Tópicos Neuronales Dinámicos han evolucionado considerablemente, incorporando técnicas avanzadas como embeddings, redes neuronales y mecanismos de autoatención para superar las limitaciones de los enfoques tradicionales. Las innovaciones recientes, como el uso de estructuras en red, modelos alineados, mecanismos de desacoplamiento y enfoques sin cadenas de Markov, 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.

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

Referencias

- Adji B. Dieng Francisco J. R. Ruiz, David M. Blei (2019). *The Dynamic Embedded Topic Model*. arXiv: [1907.05545 \[cs.CL\]](https://arxiv.org/abs/1907.05545). URL: <https://arxiv.org/abs/1907.05545>.
- Blei and Lafferty (2006). “Dynamic topic models”. In: ACM Press. DOI: [10.1145/1143844.1143859](https://doi.org/10.1145/1143844.1143859).
- Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe (Apr. 2017). “Variational Inference: A Review for Statisticians”. In: *Journal of the American Statistical Association* 112.518, 859–877. ISSN: 1537-274X. DOI: [10.1080/01621459.2017.1285773](https://doi.org/10.1080/01621459.2017.1285773). URL: <http://dx.doi.org/10.1080/01621459.2017.1285773>.
- Blei, David M., A. Ng, and Michael I. Jordan (2003). “Latent Dirichlet Allocation”. In: URL: <https://api.semanticscholar.org/CorpusID:3177797>.
- Cvejoski, Kostadin, Ramsés J. Sánchez, and César Ojeda (2023). *Neural Dynamic Focused Topic Model*. arXiv: [2301.10988 \[cs.CL\]](https://arxiv.org/abs/2301.10988). URL: <https://arxiv.org/abs/2301.10988>.
- Dieng, Adji B., Francisco J. R. Ruiz, and David M. Blei (2019). “Topic Modeling in Embedding Spaces”. In: *Transactions of the Association for Computational Linguistics* 8, pp. 439–453. URL: <https://api.semanticscholar.org/CorpusID:195886143>.
- Gershman, Samuel J. and Noah D. Goodman (2014). “Amortized Inference in Probabilistic Reasoning”. In: *Cognitive Science* 36. URL: <https://api.semanticscholar.org/CorpusID:924780>.

- Hochreiter, Sepp and Jürgen Schmidhuber (1997). “Long Short-Term Memory”. In: *Neural Computation* 9, pp. 1735–1780. URL: <https://api.semanticscholar.org/CorpusID:1915014>.
- Hoffman, Matt et al. (2013). *Stochastic Variational Inference*. arXiv: 1206.7051 [stat.ML]. URL: <https://arxiv.org/abs/1206.7051>.
- Jordan, Michael I. et al. (1999). “An Introduction to Variational Methods for Graphical Models”. In: *Machine Learning* 37, pp. 183–233. URL: <https://api.semanticscholar.org/CorpusID:2073260>.
- Kingma, Diederik P and Max Welling (2022). *Auto-Encoding Variational Bayes*. arXiv: 1312.6114 [stat.ML]. URL: <https://arxiv.org/abs/1312.6114>.
- Lin, Zhouhan et al. (2017). *A Structured Self-attentive Sentence Embedding*. arXiv: 1703.03130 [cs.CL]. URL: <https://arxiv.org/abs/1703.03130>.
- Miyamoto, Nozomu et al. (July 2023). “Dynamic Structured Neural Topic Model with Self-Attention Mechanism”. In: *Findings of the Association for Computational Linguistics: ACL 2023*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 5916–5930. DOI: 10.18653/v1/2023.findings-acl.366. URL: <https://aclanthology.org/2023.findings-acl.366>.
- Perrone, Valerio et al. (2016). *Poisson Random Fields for Dynamic Feature Models*. arXiv: 1611.07460 [stat.ML]. URL: <https://arxiv.org/abs/1611.07460>.

- Rahimi, Hamed et al. (2023). *ANTM: An Aligned Neural Topic Model for Exploring Evolving Topics*. arXiv: [2302.01501](https://arxiv.org/abs/2302.01501) [cs.IR]. URL: <https://arxiv.org/abs/2302.01501>.
- Vaswani, Ashish et al. (2023). *Attention Is All You Need*. arXiv: [1706.03762](https://arxiv.org/abs/1706.03762) [cs.CL]. URL: <https://arxiv.org/abs/1706.03762>.
- Williamson, Sinead et al. (2010). “The IBP Compound Dirichlet Process and its Application to Focused Topic Modeling”. In: *International Conference on Machine Learning*. URL: <https://api.semanticscholar.org/CorpusID:6023511>.
- Wu, Xiaobao et al. (2024). *Modeling Dynamic Topics in Chain-Free Fashion by Evolution-Tracking Contrastive Learning and Unassociated Word Exclusion*. arXiv: [2405.17957](https://arxiv.org/abs/2405.17957) [cs.CL]. URL: <https://arxiv.org/abs/2405.17957>.
- Zhang, Delvin Ce and Hady Lauw (2022). “Dynamic Topic Models for Temporal Document Networks”. In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri et al. Vol. 162. Proceedings of Machine Learning Research. PMLR, pp. 26281–26292. URL: <https://proceedings.mlr.press/v162/zhang22n.html>.

Bibliografía

- Adji B. Dieng Francisco J. R. Ruiz, David M. Blei (2019). *The Dynamic Embedded Topic Model*. arXiv: [1907.05545 \[cs.CL\]](https://arxiv.org/abs/1907.05545). URL: <https://arxiv.org/abs/1907.05545>.
- Wu, Xiaobao, Thong Nguyen, and Anh Tuan Luu (2024). “A Survey on Neural Topic Models: Methods, Applications, and Challenges”. In: *ArXiv* abs/2401.15351. URL: <https://api.semanticscholar.org/CorpusID:267297321>.
- Wu, Xiaobao et al. (2024). *Modeling Dynamic Topics in Chain-Free Fashion by Evolution-Tracking Contrastive Learning and Unassociated Word Exclusion*. arXiv: [2405 . 17957 \[cs.CL\]](https://arxiv.org/abs/2405.17957). URL: <https://arxiv.org/abs/2405.17957>.