

Reporte Final: Explorando la Detección de Temas en Conjuntos de Documentos

Carina Boyallian

December 13, 2024

1 Introducción

La detección de temas en conjuntos de documentos es un área activa en procesamiento del lenguaje natural (NLP), con aplicaciones que van desde la organización automática de textos hasta el análisis de grandes volúmenes de datos en investigación y negocios. Tradicionalmente, métodos como el Latent Dirichlet Allocation (LDA) [1] han sido utilizados para identificar temas en textos, agrupando documentos en función de tópicos comunes. No obstante, con los avances en aprendizaje automático, han surgido enfoques que combinan embeddings de palabras y documentos con modelos neuronales para mejorar la precisión y coherencia de los temas identificados. La evaluación de estos métodos y su adaptabilidad a distintos conjuntos de datos sigue siendo un desafío en la optimización de técnicas de detección de temas.

Este informe presenta los resultados de una prueba de factibilidad en detección de temas, evaluando la efectividad de cinco tipos de embeddings —TF-IDF, LSA [2], Doc2Vec [3], DistilBERT [4, 5] y LDA— en combinación con técnicas de clustering básicos y modelos avanzados de detección de temas. Para el estudio se utiliza el conjunto de datos 20Newsgroups, una referencia común en NLP por su estructura organizada en múltiples categorías temáticas.

La detección de temas ha evolucionado desde técnicas estadísticas como TF-IDF y LSA, que utilizan frecuencias de términos y descomposición en valores singulares, respectivamente, para representar palabras y temas. Aunque son efectivos, estos métodos tienen limitaciones para capturar semántica compleja. TF-IDF representa términos sin tener en cuenta el contexto más amplio, mientras que LSA permite identificar relaciones subyacentes entre palabras y temas más allá de co-ocurrencias. Aún así, ambos métodos enfrentan desafíos en la captura de relaciones semánticas profundas.

La transición hacia embeddings neuronales, como Doc2Vec y DistilBERT, ha mejorado significativamente la calidad de las representaciones de textos al capturar con mayor precisión la semántica del contexto. Doc2Vec, una extensión de Word2Vec, permite codificar documentos completos, destacándose en la identificación de similitudes semánticas al considerar el contexto global del texto. Por su parte, modelos como DistilBERT han avanzado en la representación de documentos gracias a su capacidad para aprender estructuras semánticas complejas, capturando patrones de co-ocurrencia entre palabras de manera más detallada, como se explora en [14].

El objetivo de este reporte es combinar estos embeddings con técnicas de clustering básico y dos modelos avanzados de modelado de tópicos. El primero, Embedded Topic Model (ETM) [9], es un modelo neuronal diseñado para incorporar embeddings en el análisis de tópicos. El segundo, basado en la distribución de von Mises-Fisher (vMF) [7], agrupa documentos en espacios de alta dimensionalidad utilizando distribuciones que modelan datos en la superficie de una esfera unitaria. Este enfoque ha demostrado ser efectivo para mejorar la agrupación en espacios complejos de embeddings, especialmente en casos donde existen palabras raras o stop words [8, 10, 11, 12].

2 Hipótesis Inicial

Como hipótesis inicial, esperábamos que embeddings avanzados como Doc2Vec o DistilBERT superaran a TF-IDF en términos de calidad de clustering, dado que estos modelos están diseñados para capturar relaciones semánticas profundas entre palabras y documentos. En particular, los modelos basados en transformadores, como DistilBERT, tienen la capacidad de contextualizar palabras según su entorno, lo cual debería mejorar la coherencia de los temas identificados en comparación con métodos como TF-IDF, que se basan principalmente en frecuencias y ponderaciones de términos.

3 Exploración y visualización de la base de datos 20Newsgroups

El conjunto de datos 20Newsgroups contiene aproximadamente 20,000 artículos agrupados en 20 categorías temáticas que incluyen ciencia, deportes, tecnología, religión, entre otros. Este data set se ha usado ampliamente en investigaciones de procesamiento en lenguaje natural. En la Tabla 1 hay una descripción de los 20 tópicos.

Se diseño un pipeline de limpieza que incluyó tokenización, lematatización, eliminación de stopwords, etc. Luego se analizó el balance de las 20 categorías, Fig. 1, y se hicieron las nubes de palabras por categoría, Fig. 2.

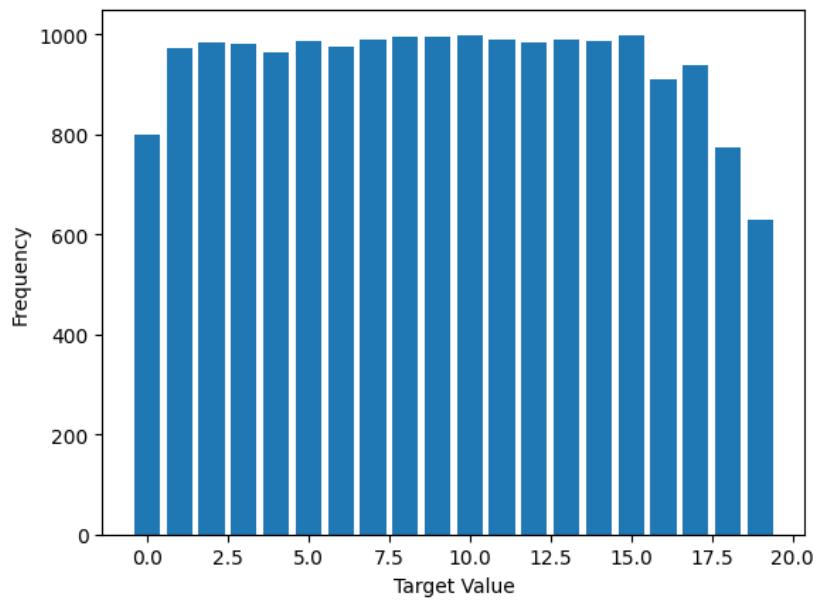


Figure 1: Descripción del balance de las clases



Figure 2: Comparación de las diferentes nubes de puntos de 4 categorías distintas.

Se visualizaron los textos a través de técnicas de reducción de dimensionalidad como t-SNE y PCA con representaciones vectoriales como TF-IDF y Word2Vec (Ver Fig. 3 y Fig. 4). Esto se puede ver en Baseline.

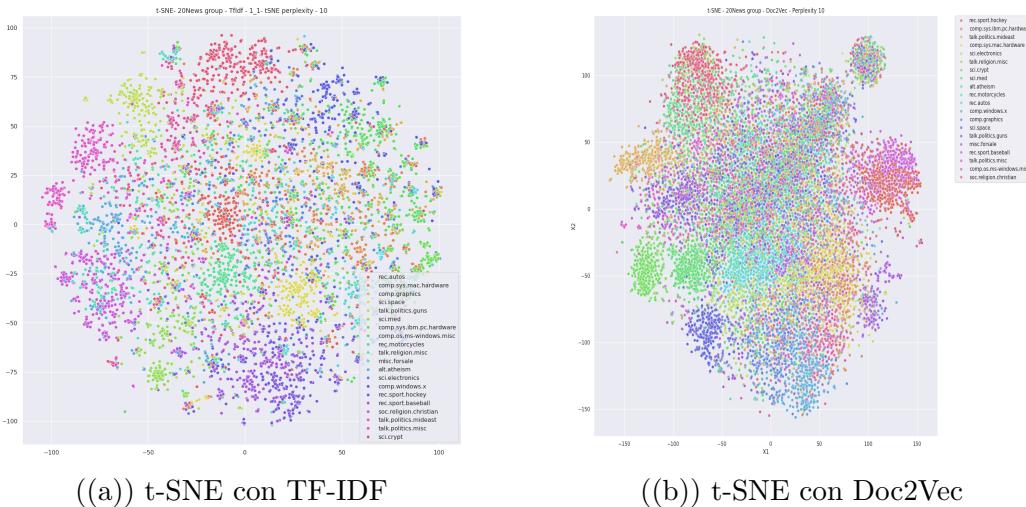


Figure 3: Comparación de t-SNE aplicado a TF-IDF y Doc2Vec.

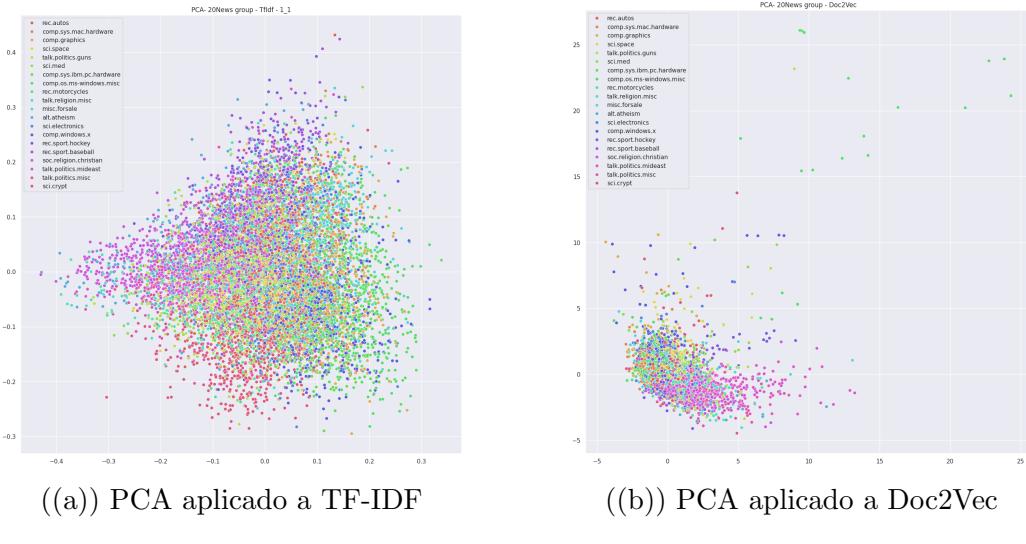


Figure 4: Comparación de PCA aplicado a TF-IDF y Doc2Vec.

La información mutua entre las categorías fue calculada para evaluar la cohesión dentro de cada clase temática.(Fig. 5)

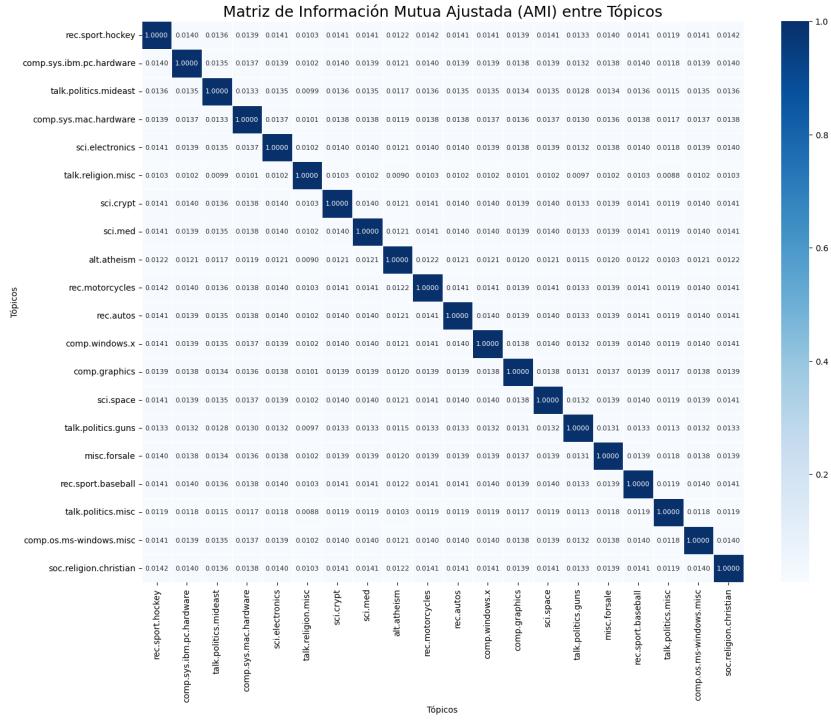


Figure 5: Resultados de la Información Mutua Ajustada (AMI) entre los tópicos originales.

4 Preliminares

A continuación presentamos una breve descripción de los modelos ETM, VMF y la métrica c_v .

ETM: Modelo de Temas Embobido (ETM), un modelo neuronal de documentos que combina modelos de temas tradicionales con embeddings de palabras, descubre temas interpretables incluso con vocabularios amplios que incluyen palabras raras y palabras de parada.

vMF: La distribución de von Mises-Fisher (vMF) es una distribución de probabilidad en espacios de alta dimensionalidad, utilizada principalmente para datos que viven en la superficie de una esfera unitaria. La vMF permite modelar la afinidad entre estos vectores, enfocándose en la dirección en lugar de la magnitud.

La medida de **coherencia** c_v es una métrica utilizada en el modelado de temas para evaluar la calidad de los temas generados. Esta métrica es ampliamente utilizada porque intenta reflejar qué tan interpretables o coherentes son los temas para los seres humanos, especialmente en términos de cómo los términos dentro de un mismo tema están relacionados semánticamente.

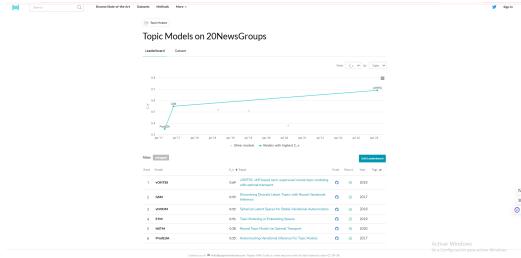


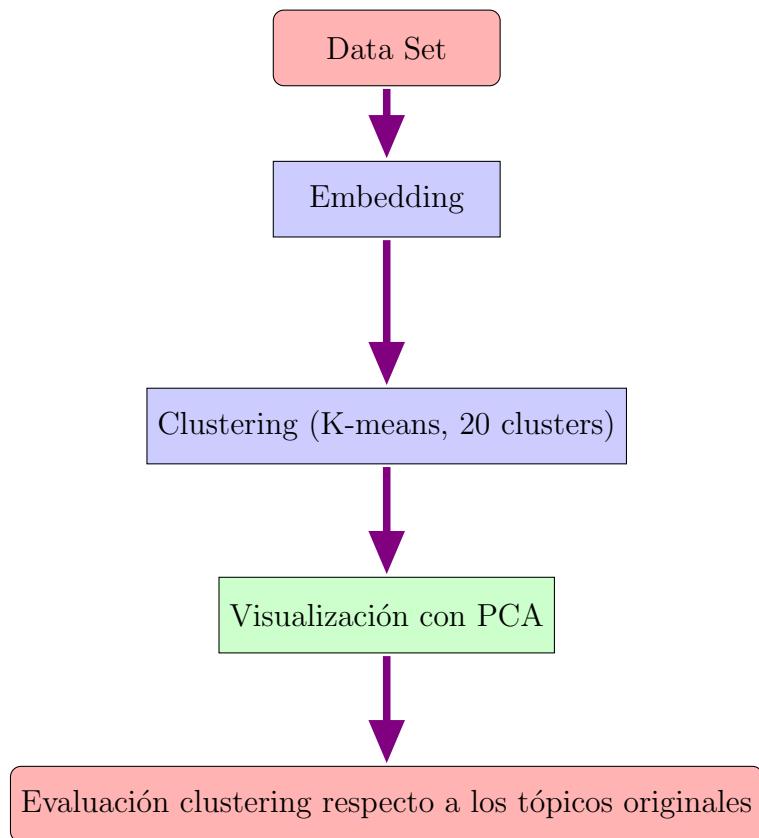
Figure 6: Benchmark 20NewsGroups

¿Qué mide c_v ? La coherencia c_v mide la consistencia semántica de los términos en cada tema. En otras palabras, evalúa si las palabras de un tema tienden a aparecer juntas en los documentos de una manera que tenga sentido y sea comprensible. Un valor de c_v más alto indica que las palabras dentro de cada tema son más consistentes y están mejor relacionadas entre sí, lo que sugiere una mayor coherencia temática.

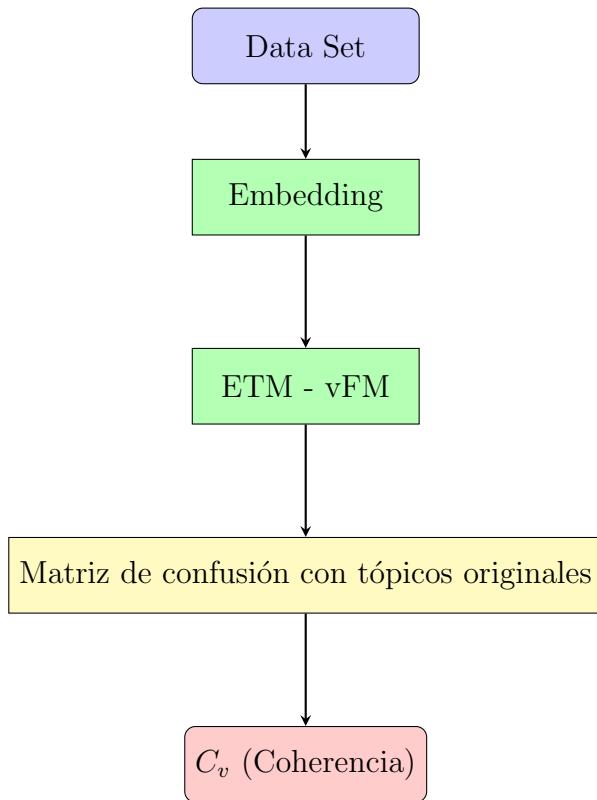
5 Metodología. Flujo de trabajo

El flujo de trabajo tuvo dos Fases, estructuradas de la siguiente manera:

FASE 1: Esta fase se desarrolló en el siguiente notebook



FASE 2. Se desarrollo en los siguientes notebooks: Doc2vec, LSA, LDA, TFIDF y DistilBert



1. **Métodos de embedding utilizados:** Se probaron cinco tipos de embeddings:

- **TF-IDF:** Un método basado en la frecuencia de términos y la inversa de la frecuencia de documentos.
- **LDA:** Latent Dirichlet Allocation, un modelo probabilístico que asigna palabras a temas.
- **LSA:** Latent Semantic Analysis, basado en descomposición en valores singulares (SVD).
- **Doc2Vec:** Modelo neuronal basado en Word2Vec que representa documentos como vectores.
- **DistilBERT:** Un modelo neuronal preentrenado basado en BERT, optimizado para eficiencia.

FASE 1:

2. **Clustering:** Se aplicó K-means con 20 clusters para cada uno de los embeddings mencionados. Para cada cluster se calcularon:

- El texto y tópico más cercano al centroide.
- El porcentaje de textos que corresponden a cada tópico original en cada cluster.
- El número de puntos en cada cluster.
- La distancia máxima y mínima al centroide.
- Pureza de los clusters respecto de los tópicos originales.
- Coherencia.

Además, se visualizó la distribución de los clusters utilizando PCA para analizar las proyecciones en los espacios generados.

3. **Matriz de confusión:** Se compararon los clusters generados vs. los tópicos originales del conjunto de datos, calculando la matriz de confusión para cada uno de los métodos probados.
4. También se calculó la **coherencia** c_v para el clustering. La medida de coherencia, es una métrica que combina similitud coseno indirecta con agregación. Esta métrica nos va a permitir comparar k -means con los otros dos modelos de modelado de tópicos.

FASE 2

Para cada embedding, se aplicó ETM y vMF, se calcularon la precisión y la coherencia c_v para poder comparar con el benchmark Topic Models on 20NewsGroups. (Fig. 6).

6 Procedimiento y Resultados

FASE 1:

- Para hacer una primera reducción de dimensionalidad, aplicamos PCA a cada embedding, luego calculamos la varianza acumulada al 90% para cada una y finalmente aplicamos K-means, para esta cantidad de componentes.
- Hicimos una visualización del clustering en 2 dimensiones:

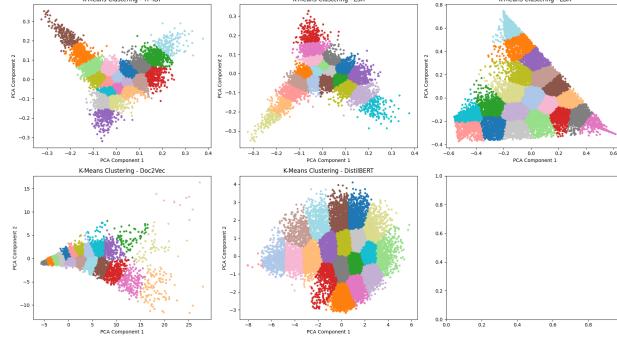


Figure 7: Clustering en 2D

Como se puede ver en la figura 7, los clusters son bastante homogeneos en todos los casos.

- Para analizar el clustering en cada caso, calculamos el tópico original mas cercano al centroide, radio de cada cluster, cantidad de puntos en cada cluster y finalmente cantidad de tópicos originales que figuran en cada cluster y el porcentaje del cluster que ocupa cada tópico original que aparece. Para visualizar este análisis, ploteamos las matrices de confusión para cada embedding:

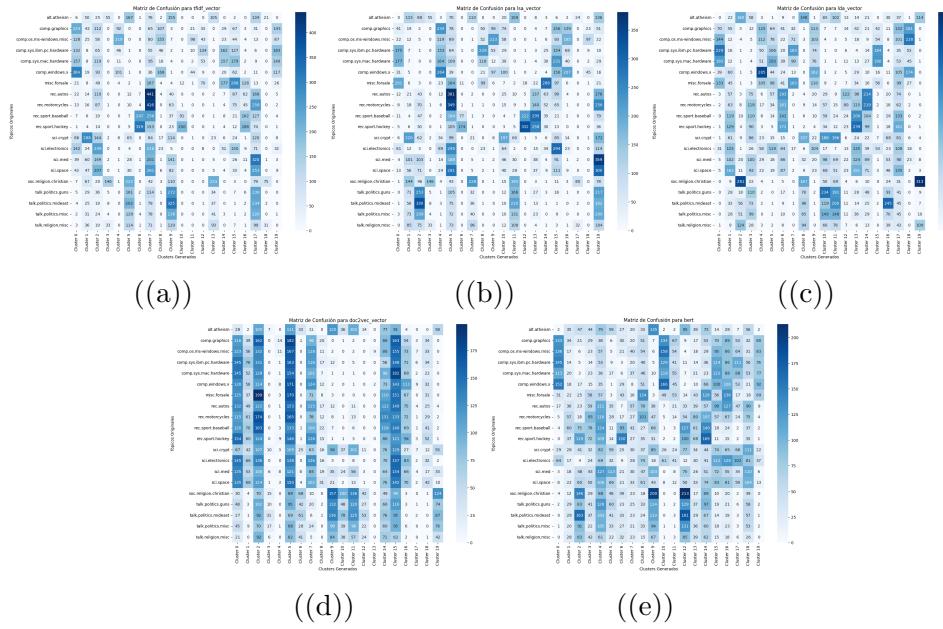


Figure 8: Matrices de confusión para K-means entre tópicos originales y clusters para cada embeddings.

Se puede ver que la pureza de los clusters no es muy buena.

- Calculamos la pureza general para cada embedding y obtuvimos los siguiente resultados:

Método	Pureza
TF-IDF	0.7912
LDA	0.2800
LSA	0.5300
Doc2Vec	0.4700
BERT	0.2100

Table 2: Comparación de Pureza

Notemos que LDA y Bert, tuvieron un desempeño malo, Doc2Vec y LSA uno medio y el mejor fue TF-IDF.

- Calculamos la coherencia C_v para cada embedding.

Método de Embedding	Coherencia C_v
K-means con TF-IDF	0.3840
K-means con LSA	0.3688
K-means con LDA	0.3557
K-means con Doc2Vec	0.4346
K-means con BERT	0.3711

Table 3: Coherencia C_v de los clusters generados en relación con los tópicos originales para cada método de embedding

En cuanto a coherencia, todos los imbeddings tienen un desempeño parecido. Se Destaca Doc2vec en este caso.

FASE 2

(a) Modelado con vMF

- Para cada uno de los 5 embeddings, implementamos el vMF dándole como conjunto de palabras para que asigne topicos, los topicos originales.
- Calculamos la C_v para cada embedding y obtuvimos los siguientes resultados:

Modelo	Puntuación de Coherencia
vMF con TF-IDF	0.3548
vMF con LSA	0.4889
vMF con LDA	0.3541
vMF con Doc2vec	0.3737
vMF con BERT	0.3656

Table 4: Puntuaciones de coherencia para vMF con diferentes embeddings

Vemos que las coherencias con vMF estan en el mismo orden de las de k -means pero en este caso, LSA supera la mejor marca de k -means obtenida con Doc2Vec.

- También calculamos las matrices de confusión de los topicos originales y de los que asigna ETM.

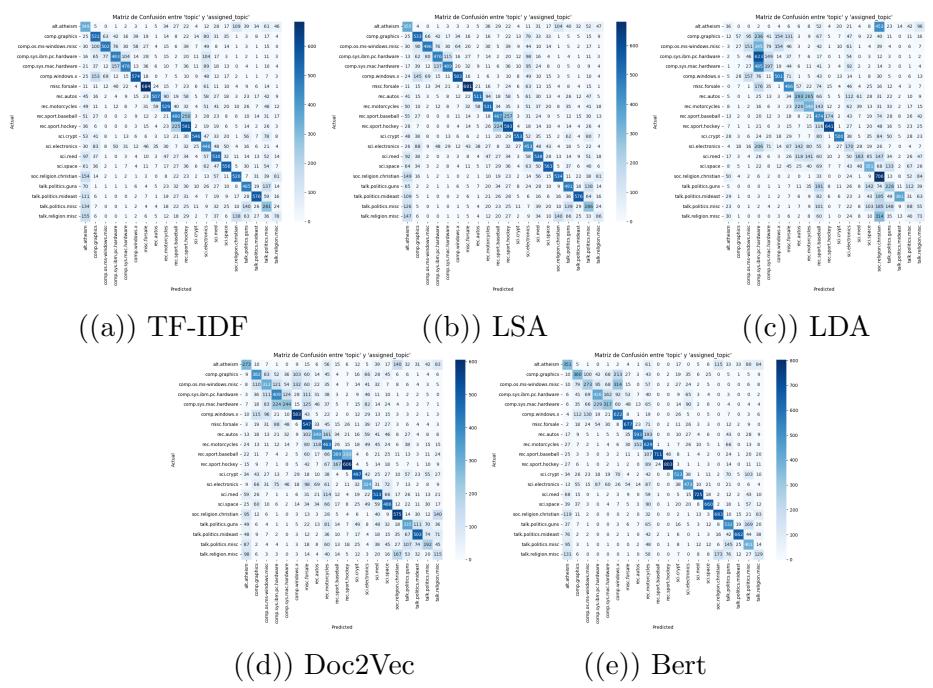


Figure 9: Matrices de confusión para vMF .

Podemos observar que con vMF las matrices de confusión mejoraron.

- Además, calculamos cual es el procentaje de textos que recibio el topico original como topico asignado por vMF para cada embedding.

Modelo	Precisión
vMF con TF-IDF	16.27%
vMF con LSA	51.97%
vMF con LDA	34.31%
vMF con Doc2vec	42.96%
vMF con BERT	55.93%

Table 5: Aciertos para vMF con diferentes embeddings

(b) Modelado con ETM

- Para cada uno de los 5 embeddings, implementamos el ETM dándole como conjunto de palabras para que asigne tópicos los originales.
- Calculamos la C_v para cada embedding y obtuvimos los siguientes resultados:

Modelo	Puntuación de Coherencia
ETM con TF-IDF	0.3701
ETM con LSA	0.3664
ETM con LDA	0.3443
ETM con Doc2vec	0.3624
ETM con BERT	0.3674

Table 6: Puntuaciones de coherencia para ETM con diferentes embeddings

- También calculamos las matrices de confusión de los tópicos originales y de los que asigna ETM. (Fig. 10)
- Además, calculamos cual es el porcentaje de textos que recibio el topico original como topico asignado por ETM para cada embedding.

Modelo	Precisión
ETM con TF-IDF	94.26%
ETM con LSA	35.49%
ETM con LDA	46.07%
ETM con Doc2vec	36.95%
ETM con BERT	61.97%

Table 7: Aciertos para ETM con diferentes embeddings

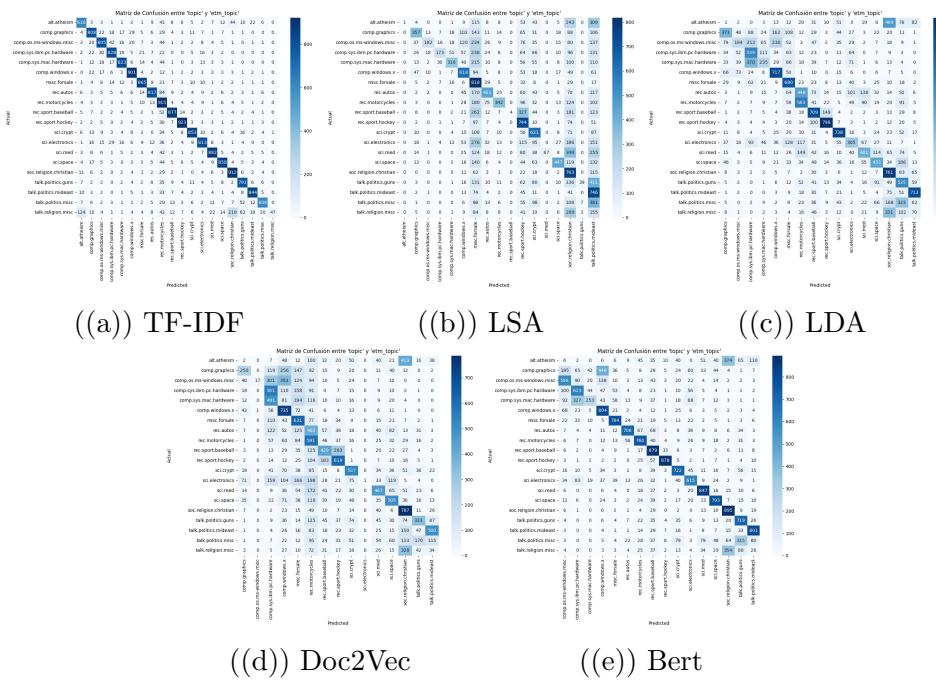


Figure 10: Matrices de confusión para ETM.

7 Resultados

El resumen de los resultados concretos obtenidos en el informe son los siguientes:

• Pureza.

TF-IDF obtuvo la mayor pureza combinada con ETM. Incluso combinado con k -means tuvo un desempeño superior a las otras combinaciones. LDA, LSA y Doc2Vec tuvieron desempeños moderados a bajos. Sin embargo, DistilBERT si obtuvo mejores resultados combinados con los modelados de tópicos que propusimos y no así con k -means.

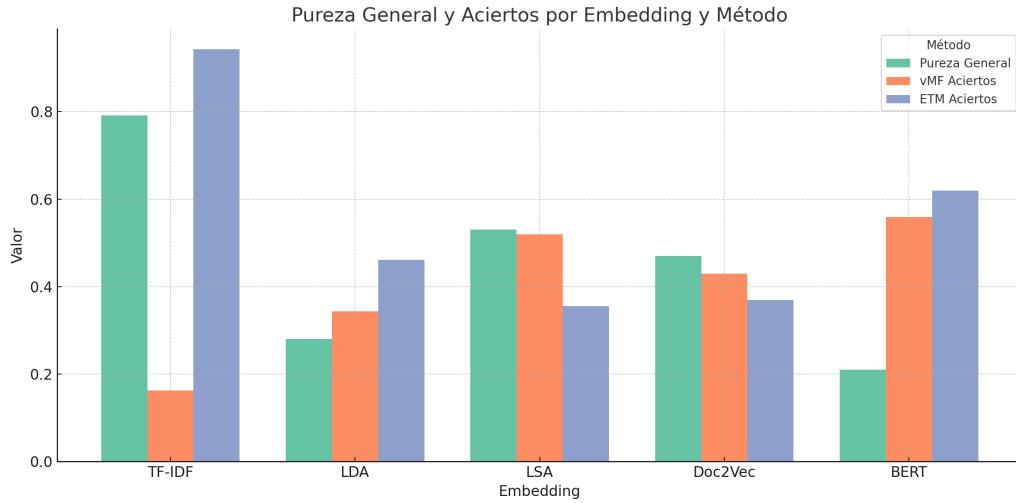


Figure 11: Pureza y aciertos por Embedding y Método de Detección de Temas

- **Coherencia Cv de los Clusters**

LSA combinado con vMF, mostró la mejor coherencia (0.4889), seguida de Doc2Vec combinada con *kmeans*. El resto de las combinaciones tuvo un desempeño parejo.

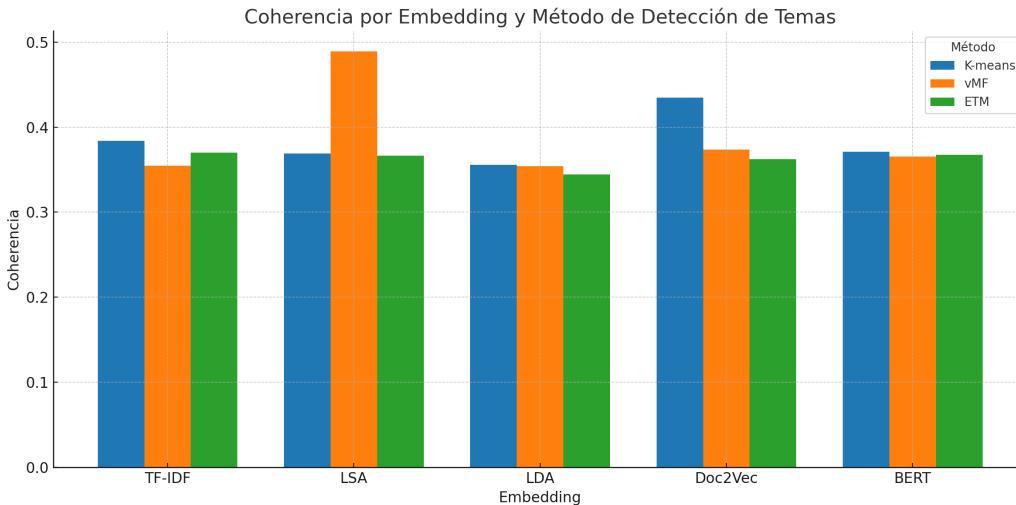


Figure 12: Coherencia por Embedding y Método de Detección de Temas

8 Conclusión

Los métodos tradicionales como TF-IDF mostraron mejor desempeño en términos de pureza en clustering básico, mientras que en coherencia, LSA lideró con vMF, seguido de Doc2Vec con k -means. Sin embargo, el rendimiento global de todos las combinaciones embedding/modelo sigue siendo bajo con respecto al Benchmarck Topic Models on 20News-Groups. Claro que esta conclusión esta limitada a los embeddings y modelos que elegimos con sus parametros básicos. Aún así, vemos que los embeddings no neuronales con modelos no neuronales tienen en ambas métricas los mejores resultados, lejos de la hipótesis inicial.

Dada esta exploración inicial, hecha con el uso de los recursos que nos da google colab, para mejorar las métricas en detección de temas, usando recursos mas potentes, se podría:

- Explorar Modelos de Clustering Alternativos: Implementar DB-SCAN y UMAP para mejorar la calidad de los clusters. Estos modelos pueden capturar estructuras de datos más complejas y reducir ruido en los resultados.
- Optimización y Ajuste Fino de Parámetros: Realizar un ajuste más detallado de los parámetros tanto de los embeddings como de K-means, como por ejemplo el número de clusters, para cada embedding específico. También ajustar los parámetros de los modelos vMF y ETM, como la regularización y el tamaño del vocabulario, para mejorar la coherencia y precisión de asignación de temas. En este trabajo usamos la arquitectura de ETM con una sola capa, se podria modificar esta arquitectura para ver si se mejoran las métricas.
- Evaluar Métricas de Coherencia Alternativas: Complementar la medida de coherencia Cv con otras métricas, como UMass o NPMI, que pueden ofrecer una evaluación adicional de la calidad de los temas.
- Integrar Embeddings Contextuales Preentrenados: Probar embeddings más avanzados o contextuales como BERT mejorado (o modelos tipo RoBERTa y ALBERT), que podrían capturar relaciones semánticas en dominios específicos mejor que los embeddings tradicionales.
- Incorporar Modelos Supervisados y Semi-Supervisados: Experimentar con modelos de aprendizaje semi-supervisado, que podrían

mejorar la agrupación al utilizar una pequeña cantidad de etiquetas, ayudando a guiar los clusters hacia temas más coherentes.

- Implementar Visualización Interactiva: Desarrollar herramientas de visualización más avanzadas (interactivas, con UMAP o t-SNE en 3D) para interpretar y analizar los resultados de los clusters y temas, lo que facilitaría la evaluación cualitativa de los modelos.
- Revisión Iterativa y Análisis de Validación Cruzada: Realizar validación cruzada para evaluar la estabilidad de los resultados con diferentes subconjuntos del conjunto de datos y garantizar la consistencia de los temas y la calidad de los clusters generados.
- Tratar de buscar en la bibliografía otros algoritmos de modelado de topicos y tratar de implementar alguna variante y/o combinación.

Estos pasos ayudarían a profundizar el análisis, evaluar mejor las fortalezas de cada método y eventualmente seleccionar la combinación de técnicas que maximicen la coherencia y la pureza en la detección de temas.

References

- [1] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). *Latent Dirichlet Allocation*. Journal of Machine Learning Research, 3, 993-1022.
- [2] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). *Indexing by latent semantic analysis*. Journal of the American Society for Information Science, 41(6), 391-407.
- [3] Le, Q., & Mikolov, T. (2014). *Distributed Representations of Sentences and Documents*. Proceedings of the 31st International Conference on Machine Learning (ICML-14), 1188-1196.
- [4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need*. Advances in Neural Information Processing Systems, 30.
- [5] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. arXiv preprint arXiv:1910.01108.
- [6] Dieng, A. B., Ruiz, F. J. R., & Blei, D. M. (2020). *Topic Modeling in Embedding Spaces*. Transactions of the Association for Computational Linguistics, 8, 439-453.

- [7] Banerjee, A., Dhillon, I. S., Ghosh, J., & Sra, S. (2005). *Clustering on the unit hypersphere using von Mises-Fisher distributions*. Journal of Machine Learning Research, 6(Sep), 1345-1382.
- [8] Xu, W., Jiang, X., Rao, S. S. H., Iannacci, F., & Zhao, J. (2023). *vONTSS: vMF based semi-supervised neural topic modeling with optimal transport*. Findings of the Association for Computational Linguistics: ACL 2023. <https://arxiv.org/abs/2305.09892>
- [9] Dieng, A. B., Ruiz, F. J. R., & Blei, D. M. (2020). *Topic Modeling in Embedding Spaces*. Transactions of the Association for Computational Linguistics, 8, 439-453. <https://arxiv.org/abs/1907.04907>
- [10] Miao, Y., Grefenstette, E., & Blunsom, P. (2017). *Discovering Discrete Latent Topics with Neural Variational Inference*. Proceedings of the 34th International Conference on Machine Learning (ICML-17), 2410-2419. <https://arxiv.org/abs/1706.00359>
- [11] Gupta, A., & Blei, D. M. (2018). *Spherical Latent Spaces for Stable Variational Autoencoders*. Advances in Neural Information Processing Systems, 31, 9068-9078. <https://arxiv.org/abs/1808.10805>
- [12] Wang, J., & Zhang, X.-L. (2019). *Deep topic modeling by multi-layer bootstrap network and lasso*. arXiv preprint arXiv:1910.10953. <https://arxiv.org/abs/1910.10953>
- [13] Panwar, M., Shailabh, S., Aggarwal, M., & Krishnamurthy, B. (2020). *TAN-NTM: Topic Attention Networks for Neural Topic Modeling*. arXiv preprint arXiv:2012.01524. <https://arxiv.org/abs/2012.01524>
- [14] C. Li, Z. Allen-Zhu, and Y. Wang, *Understanding Transformers as Learning Directed Graphs*, in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, vol. 202, pp. 25139–25167, 2023. Available: <https://proceedings.mlr.press/v202/li23p/li23p.pdf>.

Tópico	Descripción
comp.graphics	Discusiones sobre gráficos por computadora, hardware gráfico, software y programación gráfica.
comp.os.ms-windows.misc	Temas generales sobre el sistema operativo Windows de Microsoft, incluyendo soporte y problemas comunes.
comp.sys.ibm.pc.hardware	Conversaciones sobre el hardware de computadoras personales IBM y compatibles, problemas de hardware y mejoras.
comp.sys.mac.hardware	Discusiones sobre hardware de computadoras Apple Macintosh, incluyendo soporte y problemas específicos.
comp.windows.x	Temas relacionados con el sistema de ventanas X Window System, desarrollo y configuración.
rec.autos	Conversaciones sobre automóviles, temas de mantenimiento, modelos de coches y experiencias de los usuarios.
rec.motorcycles	Temas sobre motocicletas, incluyendo tipos de motos, mantenimiento y experiencias de viaje.
rec.sport.baseball	Discusiones sobre béisbol, equipos, jugadores y acontecimientos recientes en el deporte.
rec.sport.hockey	Conversaciones sobre hockey, equipos, jugadores y partidos recientes.
sci.crypt	Discusiones sobre criptografía, métodos de encriptación y seguridad informática.
sci.electronics	Temas sobre electrónica, circuitos, componentes y proyectos relacionados con la electrónica.
sci.med	Conversaciones sobre medicina, problemas de salud, tratamientos y temas de la industria médica.
sci.space	Discusiones sobre el espacio, astronomía, exploración espacial y avances en ciencia espacial.
misc.forsale	Anuncios de artículos en venta, discusiones sobre precios y consejos para compradores y vendedores.
talk.politics.misc	Temas variados sobre política, actualidad y debates generales sobre temas políticos.
talk.politics.guns	Discusiones sobre armas de fuego, legislación y temas relacionados con la posesión de armas.
talk.politics.mideast	Conversaciones sobre la situación política en el Medio Oriente, conflictos y diplomacia.
talk.religion.misc	Temas varios sobre religión, creencias y debates entre diferentes religiones y filosofías.
alt.atheism	Discusiones sobre ateísmo, debates sobre religión y puntos de vista sobre la no creencia en deidades.
soc.religion.christian	Conversaciones y debates sobre el cristianismo, ²⁰ enseñanzas, prácticas y comunidad cristiana.

Table 1: Lista de los 20 tópicos de 20 Newsgroups con descripciones en español.