Proyecto Final Modulo 2 - Aplicaciones de aprendizaje automatico de maquina

Reporte - Modelos Temáticos Supervisados vs. Embeddings Semánticos: Un Estudio sobre Clasificación de Géneros con sLDA y MPNet.

Oswald Morales[1,†], **Sergio Guarin**[1,†] **and Jose Agudelo**[2,†]

[1]*Estudiante Msc. Ingeniería, Ciencia y Tecnología*
[2]*Estudiante Msc. Matematicas Aplicadas y Ciencias de la Computacion*
[†]*Universidad del Rosario*

Junio 07, 2025

Abstract

Este trabajo explora y compara dos enfoques distintos para la clasificación automática de sinopsis cinematográficas por género: modelos supervisados de temas (sLDA) y embeddings densos generados mediante MPNet. A partir de un corpus de más de 13,000 sinopsis en inglés, se aplicó un preprocesamiento cuidadoso que incluyó la tokenización por bloques y la vectorización con Bag of Words. En el caso de sLDA, se entrenó un modelo binario por cada género para capturar las distribuciones temáticas asociadas, utilizando posteriormente regresión logística para la predicción. En contraste, el enfoque con MPNet se basó en representaciones semánticas densas de las sinopsis, evaluadas mediante un clasificador multiclase. La validación cruzada con métricas como ROC-AUC, F1-score y exactitud evidenció que, si bien MPNet ofrece resultados significativamente más robustos y generalizables, sLDA presenta limitaciones importantes derivadas del desbalanceo de clases, generando un sesgo hacia las categorías mayoritarias. Las visualizaciones de matrices de confusión y pruebas con sinopsis ficticias respaldan esta conclusión. El estudio destaca la utilidad de los modelos sLDA como herramienta interpretativa, pero resalta que su capacidad predictiva se ve comprometida en contextos con alta desproporción de clases.

Keywords: *Embeddings, AutoEncoders, Tokenizacion, Machine Learning, sLDA*

1. Introducción

En los últimos años, el aprendizaje automático ha demostrado ser una herramienta poderosa para abordar problemas complejos relacionados con el procesamiento de lenguaje natural (PLN), particularmente en tareas de clasificación de texto. Este proyecto se enmarca en dicha línea de trabajo, con el objetivo de predecir el género cinematográfico de una película a partir de su sinopsis textual, utilizando un enfoque de modelado probabilístico supervisado.

A diferencia del módulo anterior, donde se emplearon representaciones semánticas densas (embeddings) generadas por modelos preentrenados como MPNet, en esta etapa se explora una alternativa estadística interpretable: el modelo *Supervised Latent Dirichlet Allocation* (sLDA). Esta técnica extiende el modelo clásico de tópicos LDA al incorporar una variable respuesta supervisada, permitiendo capturar de manera conjunta las estructuras latentes del texto y su relación con una etiqueta de clase.

El conjunto de datos utilizado es el mismo que en el módulo anterior: *The Movies Dataset*, el cual contiene metadatos de más de 45.000 películas. A partir de este corpus, se seleccionaron únicamente aquellas películas con una sinopsis válida y un único género, acotado a trece categorías representativas. La variable `overview` se mantuvo como base textual para el modelado.

Inicialmente, se entrenó un modelo sLDA binario enfocado en un único género como clase positiva, considerando el resto de géneros como clase negativa. Este enfoque permitió explorar la relación entre los temas latentes y la probabilidad de pertenecer al género objetivo, evaluando el desempeño mediante validación cruzada y métricas como el AUC-ROC.

Posteriormente, con el fin de ampliar el análisis y establecer una comparación sistemática con los modelos basados en embeddings, se desarrolló un sistema de clasificación multiclase mediante la creación de trece modelos sLDA independientes, cada uno entrenado para identificar un género específico. Esta estrategia permitió comparar su precisión, sesgo y capacidad de generalización frente al pipeline de embeddings, identificando fortalezas y limitaciones propias del modelado temático supervisado.

2. Descripción del Dataset

El conjunto de datos utilizado en este proyecto es *The Movies Dataset*, una base de datos recopilada por MovieLens y publicada en la plataforma Kaggle. Este dataset incluye metadatos detallados de 45.466 películas estrenadas hasta julio de 2017. Entre las variables más relevantes se encuentran el título de la película (`title`), su sinopsis (`overview`), los géneros asociados (`genres`), información del elenco y equipo de producción, presupuesto estimado, ingresos generados, fechas de estreno, idiomas y métricas de popularidad como número de votos y promedio de calificación en TMDB.

Para este proyecto, se seleccionaron únicamente tres columnas: `title`, `overview` y `genres`. Posteriormente, se aplicaron una serie de filtros para garantizar la calidad y uniformidad de los datos, incluyendo:

- Eliminación de registros con sinopsis vacías.
- Conversión del campo `genres` desde formato JSON a una lista legible.
- Conservación de películas con exactamente un solo género.
- Filtrado por trece géneros predefinidos con representación suficiente: *Drama, Comedy, Documentary, Horror, Thriller, Western, Action, Animation, Science Fiction, Crime, Music, Adventure*.
- Eliminación de sinopsis que, al ser tokenizadas, se fragmentan en más de un bloque.

Luego del preprocesamiento, se obtuvo un conjunto final compuesto por 13.744 películas, cada una representada por su título, sinopsis, el género correspondiente y su codificación numérica.

3. Metodología

La tarea de clasificación se abordó como un problema supervisado binario, tomando como entrada la sinopsis textual de cada película y como salida una etiqueta que indica si la película pertenece o no al género objetivo. A diferencia del enfoque basado en embeddings utilizado en el módulo anterior, en este caso se implementó un modelo de tópicos supervisado (*Supervised Latent Dirichlet Allocation*, sLDA), el cual permite extraer temas latentes a partir del texto mientras aprende a predecir una variable de salida.

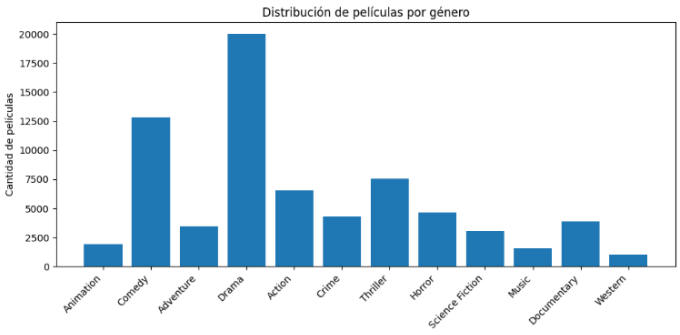

Figure 1. Distribucion de peliculas por genero

3.1. Preprocesamiento y filtrado

El proceso de preprocesamiento fue análogo al del módulo anterior, partiendo del dataset original *The Movies Dataset*. Se conservaron únicamente las columnas `title`, `overview` y `genres`. A continuación, se aplicaron los siguientes filtros:

- Eliminación de registros con sinopsis vacías.
- Conversión del campo `genres` desde formato JSON a listas legibles.
- Conservación de películas con exactamente un solo género, restringido a trece categorías válidas: *Drama, Comedy, Documentary, Horror, Thriller, Western, Action, Animation, Science Fiction, Crime, Music, Adventure*.
- Eliminación de duplicados basados en el par `(title, overview)`.

2.2 Representación del texto

Antes de aplicar el modelo sLDA, fue necesario asegurar que las sinopsis pudieran ser procesadas de forma homogénea. Para ello, se utilizó el tokenizador `SentenceTransformersTokenTextSplitter`, configurado con un máximo de 315 tokens por bloque, con el objetivo de conservar únicamente aquellas sinopsis que pudieran representarse como un único segmento de texto. Esta restricción busca garantizar que la unidad semántica de cada sinopsis no se vea fragmentada.

```
from langchain.text_splitter import
    SentenceTransformersTokenTextSplitter

token_splitter =
    SentenceTransformersTokenTextSplitter(
    chunk_overlap=0,
    tokens_per_chunk=315
)

def count_chunks(text):
    chunks = token_splitter.split_text(text)
    return len(chunks)

df['num_chunks'] = df['overview'].apply(
    count_chunks)
df_clean = df[df['num_chunks'] == 1].copy()
```

Code 1. Filtrado de sinopsis con un solo bloque

Posteriormente, se construyó una variable binaria denominada `target`, donde se asignó el valor 1 a las películas pertenecientes al género *Comedy*, y 0 al resto. Esta estructura permitió adaptar el problema de clasificación a un enfoque binario supervisado, acorde con la configuración del modelo sLDA.

```
df_clean['target'] = df_clean['genre'].apply(
    lambda x: 1 if x == 'Comedy' else 0)
```

Code 2. Creación de la variable objetivo binaria

Una vez filtrado el conjunto y definida la variable objetivo, se procedió a transformar las sinopsis a una representación de bolsa de palabras (*bag-of-words*). Esta transformación convierte cada sinopsis en un vector de ocurrencias de términos, eliminando palabras vacías en inglés y limitando el vocabulario a las 5000 palabras más frecuentes, lo que permite reducir el ruido y controlar la dimensionalidad del modelo.

```
from sklearn.feature_extraction.text import
    CountVectorizer

vectorizer = CountVectorizer(
    stop_words='english',
    max_features=5000
)

X_bow = vectorizer.fit_transform(df_clean['
    overview'])
y = df_clean['target'].values
```

Code 3. Vectorización del texto con CountVectorizer

El resultado es una matriz dispersa de dimensiones $(n_documentos, n\_terminos)$, donde cada fila representa una sinopsis y cada columna un término del vocabulario. Esta matriz se utilizó como entrada para el entrenamiento del modelo sLDA.

```
print(f"Matriz BoW: {X_bow.shape[0]} documentos,
     {X_bow.shape[1]} terminos")
```

Code 4. Verificación de dimensiones

3.2. Entrenamiento del modelo sLDA

El modelo sLDA fue implementado como un pipeline secuencial: primero, se ajustó un modelo de tópicos `LatentDirichletAllocation` (LDA) sobre la matriz de bolsa de palabras (`X_bow`) para obtener representaciones temáticas; luego, estas distribuciones se utilizaron como variables de entrada para un modelo de regresión logística encargado de predecir la pertenencia al género objetivo.

Para determinar el número óptimo de tópicos, se realizó una validación cruzada manual sobre distintos valores de p, en el rango $\{10, 12, 14, 16, 18, 20\}$, utilizando un esquema de *10-fold stratified cross-validation* y la métrica ROC-AUC como criterio de evaluación.

```

n_topics_list = [10, 12, 14, 16, 18, 20]
kfold = StratifiedKFold(n_splits=10, shuffle=
    True, random_state=42)
auc_results = {}

for n_topics in n_topics_list:
    print(f"\nEvaluando modelo con {n_topics}
    t picos...")

    lda_model = LatentDirichletAllocation(
        n_components=n_topics,
        random_state=42,
        learning_method='batch'
    )

    X_topics = lda_model.fit_transform(X_bow)

    clf = LogisticRegression(
        max_iter=1000,
        solver='liblinear',
        random_state=42
    )

    scores = []
    for train_idx, test_idx in kfold.split(
    X_topics, y):
```

```
        X_train, X_test = X_topics[train_idx], X_topics[test_idx]
        y_train, y_test = y[train_idx], y[test_idx]

        clf.fit(X_train, y_train)
        y_prob = clf.predict_proba(X_test)[:, 1]
        score = roc_auc_score(y_test, y_prob)
        scores.append(score)

    mean_auc = np.mean(scores)
    auc_results[n_topics] = mean_auc
    print(f"ROC-AUC promedio para {n_topics} t picos: {mean_auc:.4f}")
```

Code 5. Evaluación de distintos valores de k con validación cruzada

Una vez identificado el valor óptimo de p (aquél con mayor AUC promedio), se entrenó nuevamente el modelo LDA sobre todo el conjunto y se ajustó el clasificador final.

```
lda_final = LatentDirichletAllocation(
    n_components=best_k, random_state=42)
X_topics_final = lda_final.fit_transform(X_bow)
```

Code 6. Entrenamiento final del modelo LDA y extracción de temas

Sobre esta representación temática se entrenó un modelo de regresión logística, que permitió aprender la relación entre la estructura latente del texto y la probabilidad de pertenecer al género objetivo.

```
clf_final = LogisticRegression(max_iter=1000)
clf_final.fit(X_topics_final, y)
```

Code 7. Entrenamiento del modelo de regresión sobre temas

Esta combinación de modelado temático no supervisado seguido de clasificación supervisada permitió evaluar la capacidad del modelo sLDA para capturar patrones relevantes asociados al género de las películas.

4. Clasificación y Resultados

En esta sección se presentan los resultados obtenidos para el modelo sLDA entrenado con el género *Comedy* como clase positiva.

4.1. Interpretación del vector de regresión

Una de las principales ventajas del modelo sLDA es su interpretabilidad. En particular, el modelo de regresión logística ajustado sobre la representación temática permite identificar qué tópicos están más fuertemente asociados a la clase positiva. A continuación se listan los coeficientes correspondientes a cada tema, ordenados por su número interno:

Tema	Coeficiente	Impacto
Tema 6	3.5170	Muy positivo ↑
Tema 7	1.0881	Positivo ↑
Tema 4	0.8670	Positivo ↑
Tema 13	0.8596	Positivo ↑
Tema 11	0.7141	Positivo ↑
Tema 9	0.4362	Positivo ↑
Tema 1	0.3121	Positivo ↑
Tema 12	0.2547	Leve positivo ↑
Tema 16	0.0547	Neutro ↑
Tema 3	-0.0042	Neutro ↓
Tema 15	-0.5377	Negativo ↓
Tema 5	-0.6015	Negativo ↓
Tema 2	-1.2609	Negativo ↓
Tema 10	-1.2946	Negativo ↓
Tema 8	-1.9289	Muy negativo ↓
Tema 14	-3.4963	Extremadamente negativo ↓

Como puede observarse, los temas 6, 7 y 4 son los más influyentes en la predicción de comedia, mientras que los temas 14, 8 y 10 están fuertemente asociados con la clase negativa. Esta información puede ser utilizada para interpretar qué tipos de contenido temático se asocian con mayor o menor probabilidad al género de comedia.

4.2. Visualización de tópicos por clase

Para ilustrar cómo se distribuyen los temas en las distintas clases, se construyó un mapa de calor (heatmap) que muestra las frecuencias promedio de cada tópico para las observaciones positivas y negativas:

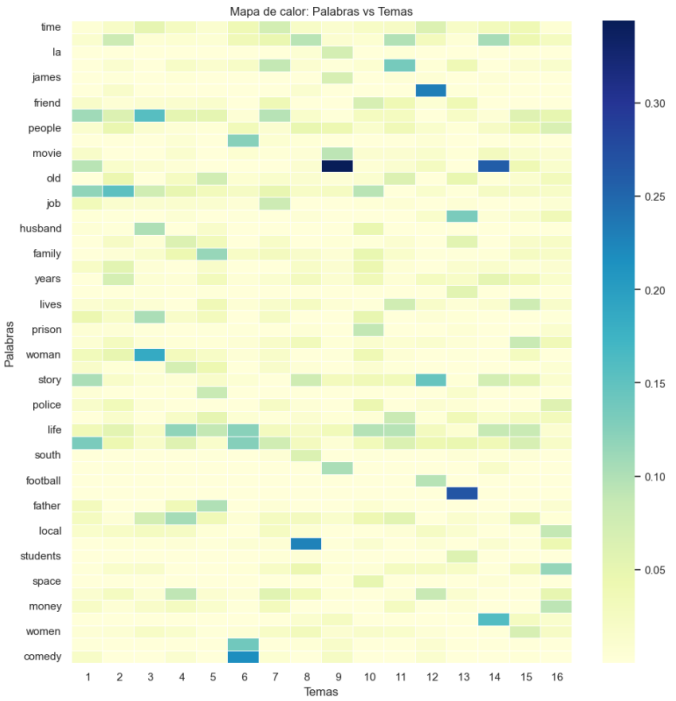

Figure 2. Mapa de calor: distribución de las 25 palabras más relevantes por tema. Cada celda representa la probabilidad de que una palabra pertenezca a un tema específico

El mapa de calor de la Figura 2 muestra la distribución relativa de las palabras más frecuentes asociadas a cada uno de los 16 temas descubiertos por el modelo LDA. Se observan agrupaciones claras, como la fuerte asociación de la palabra *movie* al Tema 9, *students* al Tema 8, o *comedy* al Tema 6. Este tipo de visualización resulta útil para interpretar el contenido semántico predominante en cada tópico y entender cómo ciertos temas contribuyen a la clasificación del género objetivo. Los temas con mayor concentración de términos humorísticos coinciden con aquellos que recibieron coeficientes positivos en la regresión logística (por ejemplo, los Temas 6, 7 y 13), lo cual valida empíricamente su relevancia predictiva.

4.2.1. Distribución de probabilidades: análisis de incertidumbre del modelo

Para evaluar la capacidad del modelo sLDA de separar correctamente las clases en el caso del género *Comedy*, se analizó la distribución de las probabilidades predichas por la regresión logística entrenada sobre los temas. La Figura 3 muestra la curva de densidad (KDE) para las predicciones del modelo, separando los casos reales positivos (comedias reales) y negativos (no comedias).

Las curvas correspondientes a ambas clases presentan una forma general similar y un amplio solapamiento en el rango de probabilidades entre 0.1 y 0.4. Esto sugiere que el modelo sLDA agrupa correctamente las observaciones según su estructura temática, pero carece de una separación clara que permita asignar alta confianza a las predicciones. En otras palabras, el modelo captura diferencias latentes entre clases, pero mantiene las probabilidades predichas en

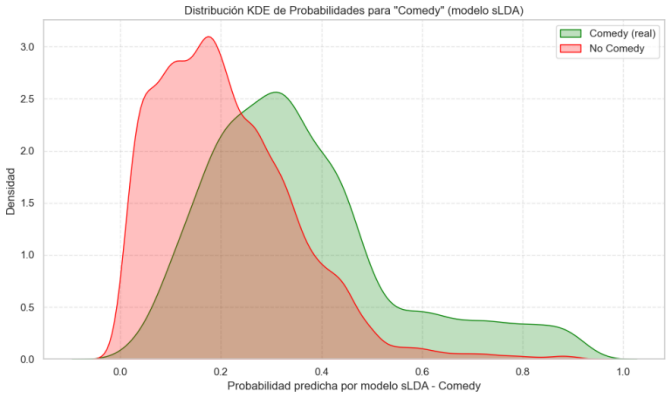

Figure 3. Distribución KDE de las probabilidades predichas por sLDA para el género *Comedy*.

un rango bajo, lo cual limita su efectividad como clasificador. Una posible mejora sería ajustar el umbral de decisión o aumentar la calibración del modelo para reflejar con mayor seguridad los casos positivos.

4.3. Evaluación con sinopsis ficticias

Como experimento adicional, se evaluó el modelo sobre un conjunto de seis sinopsis ficticias diseñadas para cubrir distintos tonos narrativos: humor absurdo, sátira política, drama romántico, realismo mágico, acción bélica y terror. Cada una fue procesada con el vectorizador entrenado y proyectada en la representación temática del modelo final.

A continuación se muestran las sinopsis y la probabilidad predicha de pertenecer al género *Comedy*:

Table 1. Probabilidades predichas por el modelo sLDA para sinopsis ficticias

#	Sinopsis	P(Comedy)
1	Two roommates accidentally adopt a raccoon, mistaking it for a cat. Chaos and laughter ensue as it destroys their apartment.	0.4805
2	In a nation where voting is determined by dance-offs, a retired ballet dancer stages a comeback to restore democracy.	0.2087
3	A young woman falls in love with a man who believes he's from another century. As love blooms, reality begins to fade.	0.1981
4	An unemployed magician opens a bakery that only sells invisible bread and becomes a global sensation.	0.1666
5	An elite soldier must infiltrate a high-security prison to retrieve stolen launch codes before time runs out.	0.2941
6	A blind sculptor starts to see through his hands and discovers terrifying secrets buried in his latest work.	0.1372

Los resultados indican que el modelo sLDA logra capturar ciertos patrones humorísticos superficiales, como se observa en la sinopsis 1 (puntaje más alto), pero presenta dificultades para detectar humor más conceptual o implícito, como en la sinopsis 4. Asimismo, tiende a clasificar incorrectamente algunos géneros con elementos fantásticos o absurdos como comedia, lo que evidencia las limitaciones del modelo basado exclusivamente en coocurrencias léxicas.

5. Discusion

Con el objetivo de contrastar el enfoque temático basado en sLDA con un modelo supervisado moderno de representación semántica, se utilizó como referencia la metodología desarrollada en el Módulo 1, la cual empleó embeddings densos generados por el modelo preentrenado `all-mpnet-base-v2`.

A diferencia del modelo sLDA, que por su naturaleza fue implementado como un conjunto de clasificadores binarios (uno por género), el enfoque basado en MPNet utilizó un único modelo multiclase. Este fue entrenado mediante la técnica *One-vs-Rest* sobre los embeddings densos generados por el modelo preentrenado. Este diseño permitió capturar simultáneamente relaciones entre géneros, así como aprovechar mejor la representación contextual profunda de las sinopsis.

5.1. Matrices de confusión

La Figura 4 y la Figura 5 muestran las matrices de confusión obtenidas para los modelos MPNet y sLDA respectivamente.

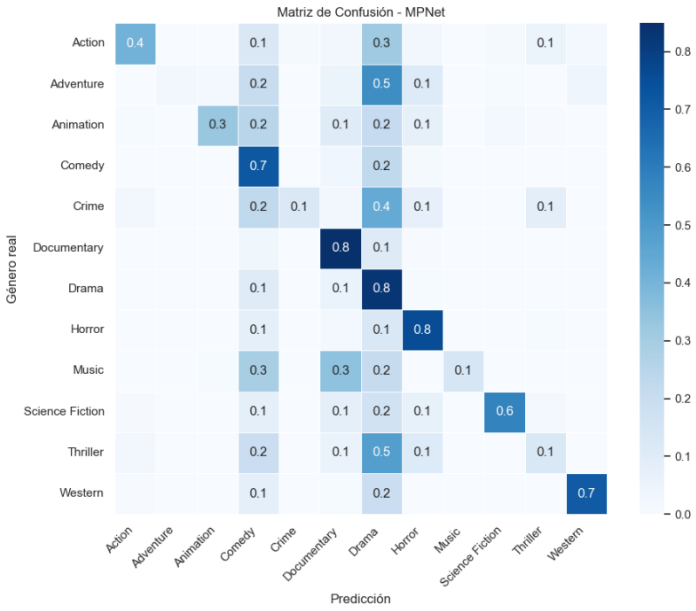

Figure 4. Matriz de confusión del modelo MPNet.

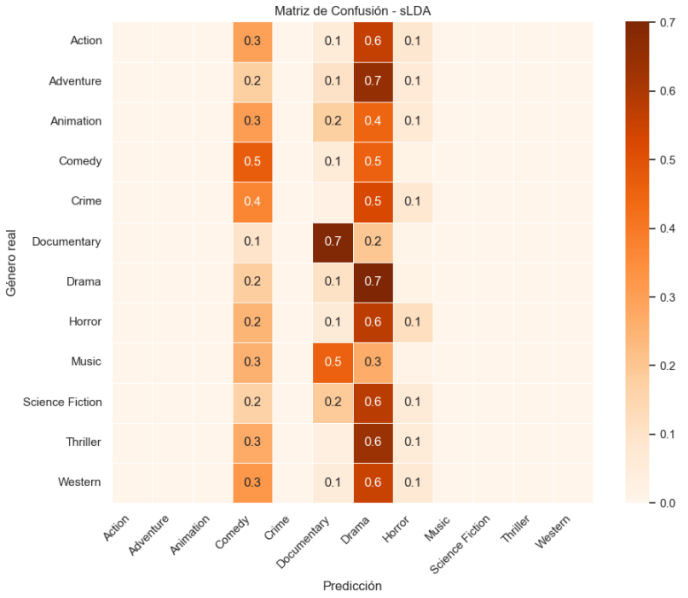

Figure 5. Matriz de confusión del modelo sLDA (multiclase agregada).

El modelo MPNet presenta una distribución de aciertos relativamente balanceada entre las clases, con valores destacables en géneros como *Comedy* (0.7), *Documentary* (0.8), *Drama* (0.8) y *Horror* (0.8). Además, los errores están distribuidos de forma más dispersa, lo que indica que el modelo no está sesgado hacia una clase específica como predicción dominante.

En contraste, el modelo sLDA muestra un comportamiento significativamente distinto. Los errores de predicción tienden a agruparse de forma vertical en unas pocas columnas (por ejemplo, *Drama*, *Documentary* y *Crime*), lo que evidencia un sesgo estructural en el modelo. Este patrón sugiere que, al enfrentarse con observaciones ambiguas o géneros menos representados, sLDA tiende a clasificarlas dentro de las clases más frecuentes del conjunto de entrenamiento.

Este fenómeno puede atribuirse en parte al desbalanceo de clases presente en el dataset. Como el modelo sLDA es entrenado de forma

binaria por género, con un enfoque temático no supervisado, los géneros minoritarios no generan suficiente señal discriminativa durante el entrenamiento, lo que provoca una concentración de probabilidades hacia las clases más dominantes.

Además, al no contar con mecanismos explícitos de ponderación de clases o ajuste dinámico del sesgo, sLDA favorece tópicos que aparecen con alta frecuencia, sin diferenciar suficientemente entre patrones comunes pero poco informativos. Por el contrario, MPNet, al codificar relaciones contextuales profundas entre palabras y frases, logra separar géneros con mayor precisión incluso en condiciones de desbalance.

Este análisis evidencia que, si bien sLDA ofrece ventajas interpretativas, su desempeño se ve afectado por su dependencia en coocurrencias superficiales y su falta de sensibilidad frente al desequilibrio de clases.

5.2. Comparación métrica por género

Además de las matrices de confusión, se evaluó el desempeño de ambos modelos por género utilizando métricas específicas como el *Recall* y el *F1-score*. Las Figuras 6 y 7 muestran estos valores para cada una de las 13 clases, permitiendo identificar diferencias relevantes en el comportamiento por categoría.

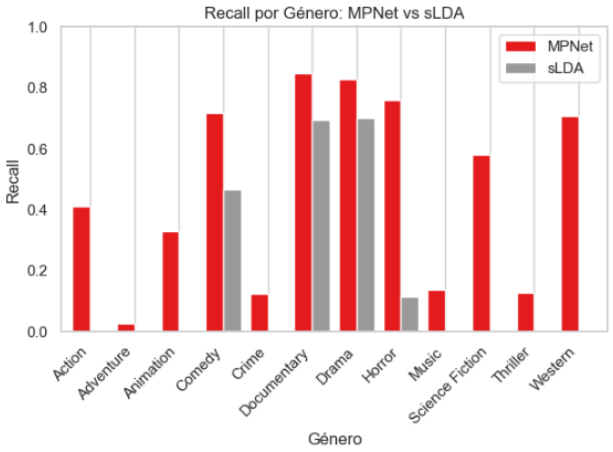

Figure 6. Recall por género: comparación entre MPNet y sLDA.

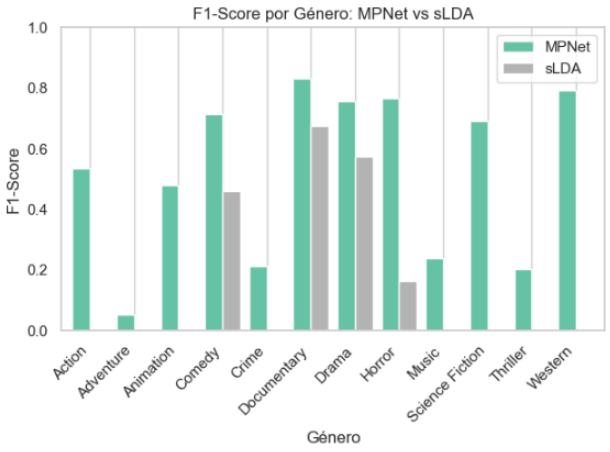

Figure 7. F1-score por género: comparación entre MPNet y sLDA.

Los resultados muestran que MPNet supera consistentemente a sLDA en la mayoría de los géneros, particularmente en aquellos con menor representación en el conjunto de datos, como *Action*, *Animation*, *Music* o *Western*. En contraste, sLDA presenta valores competitivos únicamente en géneros con mayor soporte, como *Drama* o *Documentary*, aunque incluso en estos casos tiende a subestimar las clases minoritarias.

Este comportamiento confirma que MPNet, al basarse en embeddings contextualizados, logra capturar información semántica útil para separar géneros incluso cuando los patrones léxicos son ambiguos o dispersos. Por su parte, sLDA muestra una fuerte dependencia en la frecuencia de tópicos, lo cual penaliza su rendimiento cuando no existen temas exclusivos claramente diferenciables para ciertos géneros.

4.3 Relación entre tamaño de clase y rendimiento

Para profundizar en el impacto del desbalanceo de clases, se construyó un gráfico de dispersión que relaciona el número de películas por género con el *Recall* obtenido por cada modelo. La Figura 8 muestra que MPNet mantiene un desempeño relativamente estable incluso en clases pequeñas, mientras que sLDA colapsa completamente cuando el soporte es bajo.

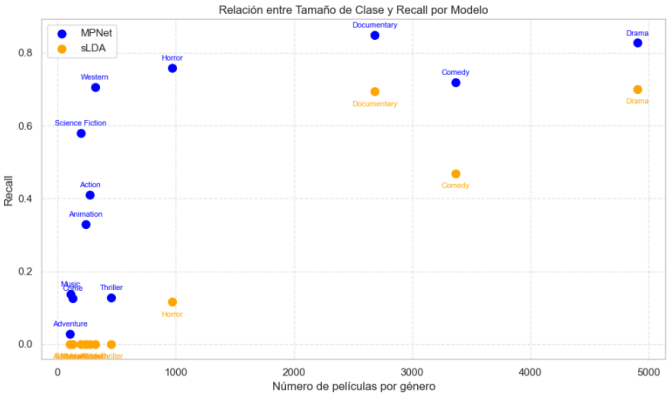

Figure 8. Relación entre tamaño de clase (número de películas) y Recall por modelo.

Esta figura refuerza la hipótesis de que sLDA es especialmente vulnerable al desbalance de clases. La representación temática del texto, al estar basada únicamente en coocurrencias, requiere una masa crítica de observaciones para aprender asociaciones robustas. En su ausencia, los modelos tienden a sesgarse hacia clases mayoritarias, como se evidenció en las matrices de confusión.

En contraste, MPNet demuestra mayor robustez al tamaño de clase, lo cual lo posiciona como una mejor opción

5.3. Análisis de exactitud y confirmación del sesgo estructural

Además de las métricas por clase discutidas anteriormente, se incorporó una medida adicional para completar el análisis: la *exactitud* de cada modelo por género. La Tabla 2 muestra los resultados de recall, F1-score y exactitud para los modelos MPNet y sLDA, junto con el tamaño de clase correspondiente a cada género.

Table 2. Comparación de Recall, F1-score y Exactitud por género para MPNet y sLDA.

Género	Tamaño	MPNet			sLDA		
		Recall	F1	Exact.	Recall	F1	Exact.
Action	275	0.411	0.537	0.986	0.000	0.000	0.980
Adventure	107	0.028	0.054	0.992	0.000	0.000	0.992
Animation	237	0.329	0.480	0.988	0.000	0.000	0.983
Comedy	3365	0.719	0.716	0.860	0.468	0.462	0.733
Crime	127	0.126	0.212	0.991	0.000	0.000	0.993
Documentary	2683	0.849	0.831	0.932	0.694	0.677	0.871
Drama	4903	0.828	0.757	0.810	0.701	0.574	0.628
Horror	971	0.759	0.767	0.871	0.000	0.000	0.880
Music	109	0.138	0.240	0.990	0.000	0.000	0.992
Science Fiction	195	0.579	0.693	0.993	0.000	0.000	0.990
Thriller	455	0.127	0.205	0.967	0.000	0.000	0.967
Western	317	0.707	0.792	0.991	0.000	0.000	0.977

Los valores de **exactitud** en el modelo sLDA son consistentemente altos (en muchos casos superiores al 0.97), incluso cuando el *recall* y

el *F1-score* son exactamente cero. Esta paradoja indica que el modelo no está aprendiendo a distinguir correctamente las instancias de la clase positiva, sino que simplemente predice de forma sistemática la clase negativa. En otras palabras, sLDA se comporta como un modelo que dice "no" ante cualquier observación, lo cual puede resultar en una alta exactitud superficial si la clase positiva es escasa, pero a costa de un rendimiento inútil en términos de cobertura.

Este comportamiento es característico de modelos que operan sobre datos desbalanceados sin mecanismos explícitos de corrección. En el caso de sLDA, esto se traduce en una tendencia a aprender únicamente los patrones mayoritarios y a ignorar las señales débiles que podrían asociarse con clases minoritarias.

6. Conclusion

Este trabajo evaluó y comparó dos enfoques metodológicamente distintos para la clasificación automática de géneros cinematográficos a partir de sinopsis textuales. Por un lado, se utilizó un modelo de tópicos supervisado (sLDA), y por el otro, un modelo de aprendizaje supervisado sobre embeddings densos derivados de MPNet. Ambos fueron entrenados sobre el mismo conjunto de datos y evaluados con métricas comunes que incluyen *Recall*, *F1-score* y matrices de confusión multiclase.

Los resultados experimentales muestran de manera consistente que el enfoque basado en MPNet ofrece un rendimiento superior en la mayoría de los géneros, logrando una mayor cobertura en clases minoritarias y una mejor separación semántica entre géneros. Esto se debe a su capacidad para capturar relaciones contextuales profundas entre palabras, lo cual permite al modelo generalizar incluso cuando el número de ejemplos es limitado.

En contraste, sLDA presenta un comportamiento más limitado, con un rendimiento aceptable únicamente en géneros con gran representación en el conjunto de datos. La representación temática extraída por sLDA, al estar basada exclusivamente en coocurrencias de palabras bajo un esquema de bolsa de palabras, carece de sensibilidad semántica y depende fuertemente de patrones estadísticos frecuentes.

Una de las conclusiones más relevantes del análisis es que el desbalance de clases en el conjunto de datos tiene un efecto crítico sobre el rendimiento de sLDA. En particular, se observó que el modelo tiende a agrupar términos temáticamente en torno a las clases mayoritarias y a ignorar las clases minoritarias, ya que estas no generan suficiente información estructural para ser aprendidas de manera efectiva. Esto significa que, en la práctica, el modelo sLDA se entrena para decir "no" ante la clase minoritaria, lo cual puede resultar en un *recall* nulo o extremadamente bajo. El sesgo sistemático observado en las matrices de confusión —con predicciones concentradas en unas pocas clases dominantes— es una manifestación directa de este fenómeno.

A pesar de estas limitaciones, sLDA conserva una ventaja fundamental: su interpretabilidad. La posibilidad de inspeccionar directamente los tópicos generados y los coeficientes del clasificador permite entender qué contenido temático está asociado a cada clase, lo que resulta valioso en contextos donde la explicabilidad del modelo es prioritaria.

Este estudio confirma que la elección del modelo depende de los requerimientos específicos del problema. Si la prioridad es la precisión y generalización en escenarios con datos desbalanceados, modelos basados en transformadores como MPNet son altamente recomendables. En cambio, si se busca comprensión estructural, análisis exploratorio o visualización temática, sLDA ofrece un marco adecuado, siempre que se maneje cuidadosamente el balance y volumen de datos.

Cabe hacer una aclaración metodológica importante: aunque a lo largo del trabajo se hizo referencia a sLDA como marco conceptual, la implementación utilizada no corresponde al modelo sLDA original propuesto por Blei et al. (2008). En su formulación canónica, sLDA es un modelo generativo conjunto que ajusta simultáneamente la distribución de temas y la predicción de la variable objetivo dentro de un mismo proceso probabilístico. En cambio, el enfoque aplicado aquí consiste en una combinación secuencial de un modelo LDA no supervisado seguido por una regresión logística independiente. Esta aproximación es ampliamente usada en la práctica como una forma efectiva de evaluar el poder discriminativo de los temas inferidos, pero no equivale al sLDA original desde el punto de vista probabilístico.

En futuras implementaciones, combinar ambos enfoques —utilizando por ejemplo tópicos extraídos por sLDA como características adicionales en modelos modernos— podría permitir alcanzar un balance entre interpretabilidad y rendimiento predictivo.

Github Link

Repositorio del proyecto en GitHub

References

[1] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022.

[2] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv preprint arXiv:1908.10084*.

[3] Hugging Face. (2020). all-mpnet-base-v2 [Model]. Recuperado de https://huggingface.co/sentence-transformers/all-mpnet-base-v2.

[4] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

[5] Scikit-learn developers. (2025). CountVectorizer. https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html.

[6] Scikit-learn developers. (2025). StratifiedKFold. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html.

[7] Scikit-learn developers. (2025). LogisticRegression. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html.

[8] Scikit-learn developers. (2025). roc_auc_score. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html.

[9] LangChain. (2025). SentenceTransformersTokenTextSplitter. https://python.langchain.com.

[10] Mcauliffe, J. D., & Blei, D. M. (2008). Supervised topic models. *Advances in Neural Information Processing Systems*, 20.