Taller 2 – Aplicaciones de ML

- Leandro Sánchez
- Leonardo Vargas

Link Repositorio: https://github.com/LeStark/Tarea-2-MLA

Junio de 2025

---

1. Presentación del problema

Este proyecto se enfoca en el problema específico de clasificar sinopsis de películas según su género, empleando técnicas de modelado estadístico de tópicos para capturar patrones semánticos relevantes presentes en el texto.

El objetivo principal es construir un clasificador binario que determine si una sinopsis pertenece a una categoría específica, en este caso, el género Drama (y_real = 1) frente a otras categorías (y_real = 0). Para ello, se empleó el conjunto de datos *The Movies Dataset*, disponible en Kaggle, el cual proporciona metadatos de más de 45.000 películas. Se trabajó con una versión filtrada que conserva únicamente aquellas películas que presentan una sinopsis válida y un único género asociado.

El conjunto de datos contiene una variedad de atributos; el atributo central analizado en este estudio fue overview, correspondiente a la sinopsis textual de cada película, sobre el cual se aplicaron técnicas de procesamiento de lenguaje natural y modelado temático para extraer representaciones significativas.

La siguiente gráfica muestra la distribución de películas por género en el dataset procesado. Se observa una marcada predominancia del género Drama. Esta concentración motivó la elección del género Drama como clase positiva en la tarea de clasificación.

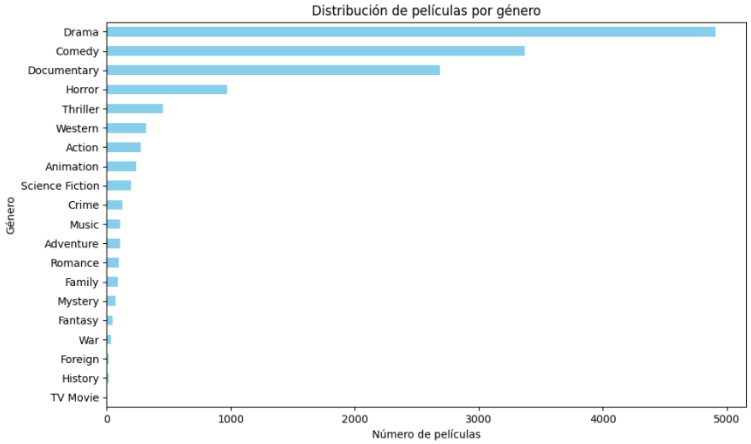

Ilustración 1Frecuencia de cada género

2. Métodos

2.1 Preprocesamiento de texto

Para preparar las sinopsis de las películas como entrada al modelo, se utilizó la librería spaCy, aplicando un proceso de limpieza y normalización lingüística que incluyó:

- Eliminación de caracteres especiales, URLs y direcciones de correo electrónico.
- Lematización de palabras para obtener sus formas base.
- Filtrado de tokens según su categoría gramatical (sustantivos, verbos y adjetivos).

- Eliminación de palabras vacías (*stopwords*) y tokens con menos de tres caracteres.

Para realizar estas tareas, se utilizó el modelo lingüístico en_core_web_sm de spaCy, que provee los recursos necesarios para el análisis morfosintáctico del inglés, incluyendo la tokenización, el etiquetado gramatical (POS tagging) y la lematización.Este preprocesamiento resultó en una lista depurada de palabras clave para cada sinopsis, facilitando una representación semántica más precisa para el modelado temático posterior.

2.2 Embeddings con sLDA

Se empleó la clase SLDAModel de la librería tomotopy en modo binario (vars='b') para generar representaciones temáticas supervisadas. Cada sinopsis se asoció con una etiqueta binaria (y_real) y se entrenó el modelo para obtener la distribución de tópicos correspondiente a cada documento.

Para identificar el número óptimo de tópicos (k), se aplicó una validación cruzada de 5 pliegues evaluando valores de k entre 10 y 20. El desempeño se midió mediante regresión logística sobre los embeddings generados, y se encontró que k = 12 ofrecía el mejor resultado promedio.

Valor de k	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Accuracy promedio
10	0,73	0,725	0,701	0,72	0,697	0,7145
11	0,731	0,725	0,708	0,723	0,697	0,7168
12	0,729	0,719	0,718	0,73	0,71	0,7211
13	0,735	0,711	0,716	0,717	0,701	0,716
14	0,721	0,715	0,716	0,72	0,716	0,7176
15	0,733	0,717	0,699	0,723	0,717	0,718
16	0,721	0,709	0,714	0,722	0,713	0,7159
17	0,745	0,718	0,718	0,731	0,711	0,7248
18	0,725	0,717	0,709	0,72	0,706	0,7154
19	0,721	0,716	0,714	0,733	0,717	0,7204
20	0,727	0,725	0,71	0,721	0,711	0,7188

Tabla 1 Resumen del accuracy alcanzado en la validación cruzada

2.3 Clasificador supervisado

Con las distribuciones temáticas obtenidas para cada sinopsis, se entrenó un modelo de regresión logística con el objetivo de predecir la variable binaria y_real. La evaluación del desempeño del clasificador se realizó en cada pliegue de validación cruzada utilizando la métrica de accuracy, que mide la proporción de predicciones correctas.

3. Análisis

El análisis de los coeficientes de regresión logística obtenidos en el modelo muestra con claridad cuáles temas están más asociados con la clase positiva o negativa (ver Tabla 2). La Ilustración 2 presenta una visualización detallada de los 17 temas descubiertos por el modelo sLDA, donde cada subgráfico muestra las diez palabras más representativas de un tema,

ordenadas según su peso (importancia relativa dentro del tema). Esta representación permite identificar con claridad la naturaleza semántica de cada grupo temático aprendido por el modelo.

Tema	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Coef.	-4,96	-0,79	-1,52	-0,04	-0,93	0,70	2,96	1,77	0,04	2,99	-1,45	-1,06	-6,08	-2,76	-4,04	-3,22	-4,78

Tabla 2 Coeficientes de la regresión por tema

Por ejemplo, el Tema #0 se centra en elementos relacionados con documentales y producción cinematográfica, mientras que el Tema #2 agrupa palabras vinculadas a crímenes y fuerzas del orden. El Tema #6, altamente correlacionado con la clase positiva (según el coeficiente de regresión), refleja narrativas de vida y superación, lo que puede explicar su asociación con el género dramático. De manera similar, el Tema #7 muestra una fuerte carga semántica vinculada a relaciones amorosas y matrimonios, también típicos del drama.

Ilustración 2 Peso relativo de las 10 palabras mas representativas por tema

En contraste, otros temas como el Tema #16, centrado en comedia y presentaciones en vivo, y el Tema #12, asociado a misterio y eventos sobrenaturales, pueden representar géneros no dramáticos, lo que se alinea con sus coeficientes de regresión negativos. Esta visualización, en conjunto con el análisis cuantitativo, refuerza la idea de que el modelo no solo captura patrones temáticos consistentes, sino que los alinea eficazmente con la tarea supervisada de clasificación.

El modelo sLDA también presenta algunas limitaciones que es importante considerar. Al ser un modelo probabilístico complejo, su rendimiento puede verse afectado por el tamaño del corpus, la calidad del preprocesamiento textual y la configuración de hiperparámetros. Además, en corpus muy diversos, los temas generados pueden ser ambiguos o poco interpretables, lo que dificulta la lectura semántica y limita su utilidad como explicadores. Pese a ello, en este caso de estudio, sLDA demostró ser una herramienta robusta para vincular la modelación de temas con tareas supervisadas, aportando tanto precisión como interpretabilidad al proceso de clasificación.