Tarea #: 4 Tema: Clasificación de datos utilizando texto Fecha entrega: 11:59 pm Junio 06 de 2025 Objetivo: Utilizar modelos de regresión logística y lstm para crear un modelo de clasificación utilizando datos reales .

Punto 2 clasificador del tipo de contenido basados en el titulo.

2.a.i - Conteo de Palabras

Se normalizan los títulos:
- Conversión a minúsculas.
- Eliminación de tildes con `unicodedata`.
- Separación por palabras usando expresiones regulares.
Luego, se construye un vocabulario de palabras únicas y se genera una matriz de conteo de palabras por título. Cada fila representa un título, y cada columna una palabra; el valor es cuántas veces aparece esa palabra en ese título.

2.a.ii - Remoción de Stopwords y Matriz TF

Se utiliza NLTK y sklearn para eliminar palabras vacías (stopwords) en español e inglés. Luego, se genera una nueva matriz llamada TF (Term Frequency), que normaliza el conteo de palabras dividiendo por el total de palabras en cada título.

2.a.iii - Cálculo del Vector IDF

Se calcula el número de títulos en los que aparece cada palabra del vocabulario y se construye el vector IDF usando la fórmula:
log((n_titulos + 1) / (n_titulos_con_palabra + 1))
Esto da más peso a palabras menos comunes.

2.a.iv - TF-IDF y Separación Train/Test

Se multiplica la matriz TF por el vector IDF para obtener la matriz TF-IDF final. Esta matriz se divide en entrenamiento y prueba usando `train_test_split`. Las etiquetas (categoría) son codificadas numéricamente con `LabelEncoder`.

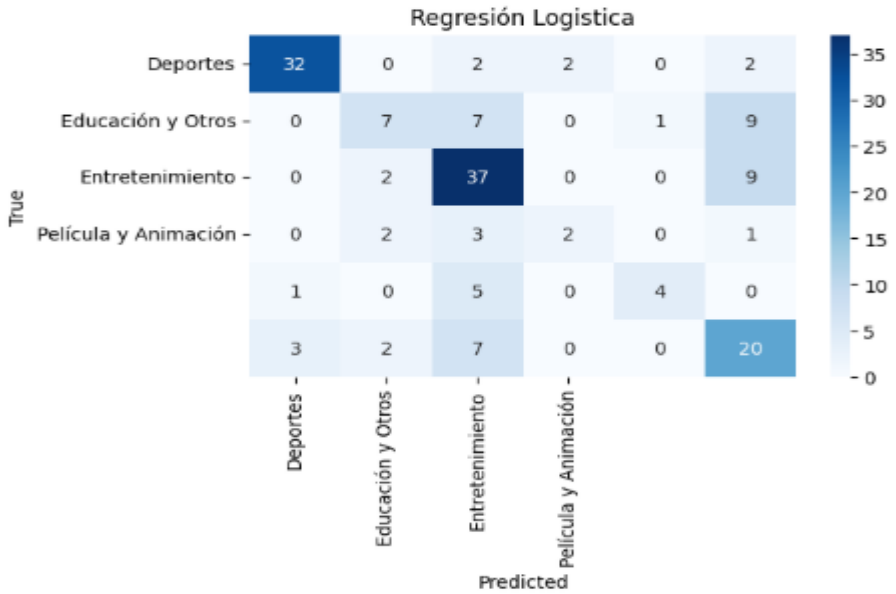
2.a.v - Entrenamiento y Evaluación de Modelos

Se entrenaron tres modelos de clasificación:
- Regresión Logística.
- Random Forest.
- LSTM (utilizando una capa de embedding).
Para los dos primeros se usó la matriz TF-IDF. Para LSTM se empleó un `Tokenizer` con `pad_sequences`.

Graficas y resultados obtenidos

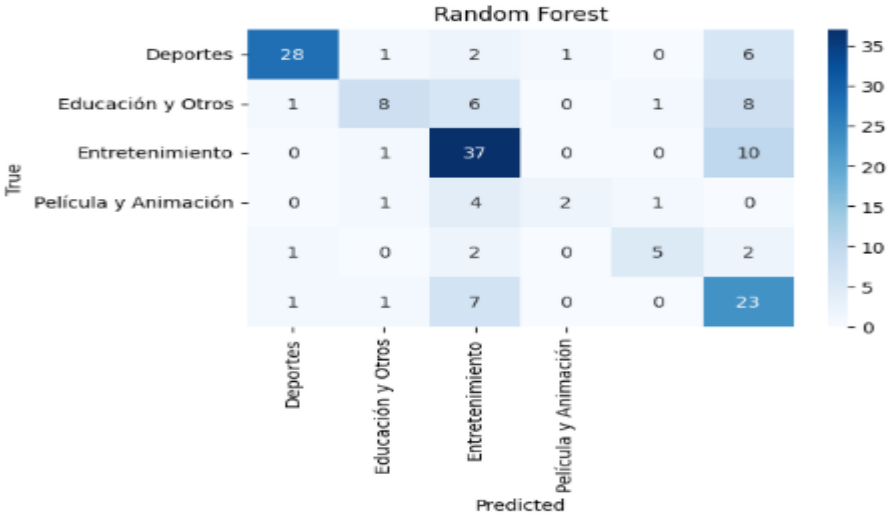
logModel

	precision	recall	f1-score	support
Deportes	0.89	0.84	0.86	38
Educación	0.54	0.29	0.38	24
Entretenimiento	0.61	0.77	0.68	48
Gente y Blogs	0.50	0.25	0.33	8
Otros	0.80	0.40	0.53	10
Película y Animación	0.49	0.62	0.55	32
accuracy			0.64	160
macro avg	0.64	0.53	0.56	160
weighted avg	0.65	0.64	0.63	160

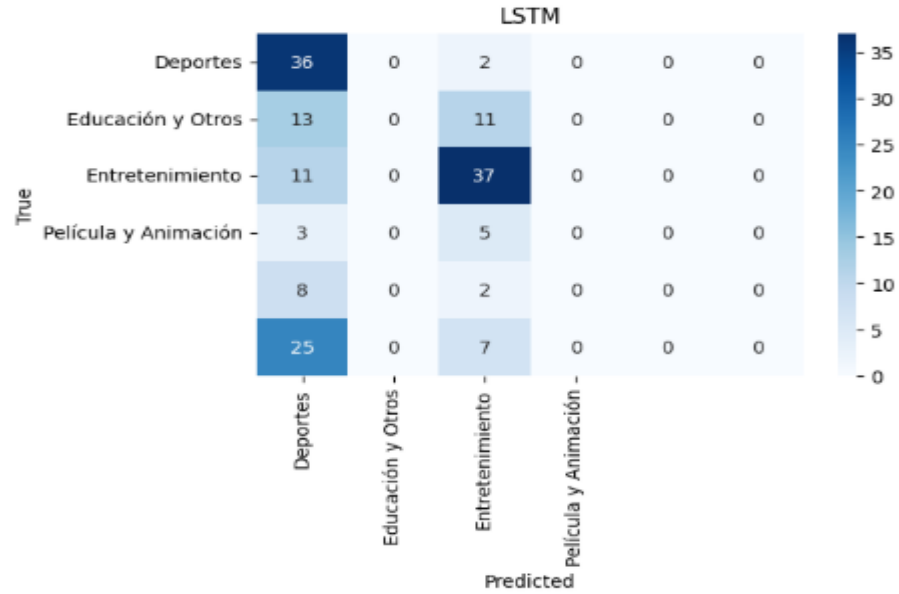


Random Forest

	precision	recall	f1-score	support
Deportes	0.90	0.74	0.81	38
Educación	0.67	0.33	0.44	24
Entretenimiento	0.64	0.77	0.70	48
Gente y Blogs	0.67	0.25	0.36	8
Otros	0.71	0.50	0.59	10
Película y Animación	0.47	0.72	0.57	32
accuracy			0.64	160
macro avg	0.68	0.55	0.58	160
weighted avg	0.68	0.64	0.64	160



LSTM



Comparación Metricas

	Modelo	Accuracy	Precision promedio	Recall promedio	F1-score promedio
0	Regresión Logística	0.63750	0.636952	0.529934	0.556126
1	Random Forest	0.64375	0.676361	0.551626	0.578987
2	LSTM	0.45625	0.158854	0.286367	0.199671

Resultados Obtenidos en Kaggle