Instituto Politécnico Nacional
Escuela Superior de Cómputo

PRÁCTICA 3: TEXT CLASSIFICATION

Natural Language Processing

Integrantes:

- **Ortega Prado Mauricio**
- **Palmerin García Diego**
- **Pérez Gómez Andres**

Grupo: 7CM2

Profesor: Joel Omar Juárez Gambino

30 DE ABRIL DE 2025

Tarea a resolver

El objetivo de esta práctica fue diseñar y evaluar un sistema de clasificación de texto, estos textos son artículos científicos provenientes de ArXiv, estos se obtuvieron anteriormente realizando web scrapping para formar un corpus de 150 documentos que entran en las secciones de "Computation and Language" y "Computer Vision and Pattern Recognition". Para realizar esta clasificación, se utilizo el titulo y resume de cada articulo como entrada textual, y se intentó predecir la categoría como salida.

Se trabajo en representaciones textuales en unigramas, utilizando tres enfoques; frecuencia binaria, frecuencia absoluta y TF-IDF. Posteriormente, se entrenaron distintos modelos de aprendizaje automático para evaluar su desempeño en esta tarea de clasificación.

Métodos de machine learning seleccionados

Se utilizaron los siguientes modelos para la clasificación de texto:

Naïve Bayes Multinomial (MultinomialNB): conocido por su eficiencia en tareas de clasificación de texto.

Regresión Logística (LogisticRegression): un clasificador lineal robusto con buen rendimiento en problemas de texto.

Máquinas de Vectores de Soporte (SVC): probadas tanto con kernel por defecto como con kernel lineal.

Perceptrón Multicapa (MLPClassifier): una red neuronal alimentada hacia adelante con diferentes capas ocultas.

Hiperparámetros ajustados

Durante los experimentos se variaron los siguientes hiperparámetros de cada modelo para evaluar su impacto:

Logistic Regression:

- max_iter: 200, 1000
- C: 0.5, 1.5 (controla regularización)

SVC:

- kernel: 'rbf', 'linear'
- C: 1, 2

MLPClassifier:

- hidden_layer_sizes: (100,), (200, 100), (300, 200, 100)
- max_iter: 300, 500

MultinomialNB:

- Sin ajuste de hiperparámetros (modelo base)

Reportes de clasificación

Machine learning method	ML method parameters	Text representation	Average f-score macro
Naive Bayes	Default	Frequency	0.9657
Logistic Regression	max_iter = 1000, C = 0.5	TF-IDF	0.9321
SVM	Kernel = rbf', C = 1.0	TF-IDF	0.9306
MLPClassifier	hidden_layer_sizes = 100, max_iter = 300	Binaria	0.9321