Machine Learning Assignment 2 Report

Text Classification with Logistic Regression and SVM on DBpedia14

Data Preprocessing

I used **5,000 samples for training** and **2,000 for testing**. The **"content"** column was used as input, and **"label"** as the target. Data was split into:

- **Training Set:** 4,000 samples
- **Development Set:** 1,000 samples
- **Test Set:** 2,000 samples
  No missing values were found.

Feature Engineering

Text was converted into numerical features using **TF-IDF Vectorization** with **3,000 features**.

Dataset	**Shape (Samples, Features)**
Train	(4000, 3000)
Dev	(1000, 3000)
Test	(2000, 3000)

Model Training & Tuning

Two models were trained:
✓ **Logistic Regression** (solver=lbfgs, max_iter=1000)
✓ **SVM** (kernel=linear, C=1.0)
Hyperparameter tuning using **GridSearchCV** found:

- **Best Logistic Regression Parameter:** C = 10
- **Best SVM Parameter:** C = 1

Model Evaluation

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**
Logistic Regression	94.2%	0.9412	0.9409	0.9409

Model	Accuracy	Precision	Recall	F1-score
SVM (Linear Kernel)	94.9%	0.9485	0.9471	0.9476

- **SVM performed slightly better than Logistic Regression.**
- **Both models showed strong classification performance across all classes.**

Documentation were also done using Sphinx and is shown as HTML:

Github repo:

https://github.com/ZohrehSamimi/AssignmentIIMachineLearning.git

Documentation:

Welcome to AssignmentII_MachineLearning’s documentation! — AssignmentII_MachineLearning 1 documentation