Digital Empowerment Network

Machine Learning

Week 05

Text Classification with NLP and Word Embeddings

Mentor: Hussain Shoaib

Text Classification with NLP and Word Embeddings

Objective:

To implement a Natural Language Processing (NLP) pipeline for text classification using different feature extraction techniques (TF-IDF, Word2Vec/Embeddings) and evaluate multiple classifiers.

Step 1: Dataset Selection

Choose one of the following text datasets:

- IMDB Movie Reviews (sentiment classification)
- SMS Spam Collection Dataset (spam vs ham)
- News Category Dataset
- Any dataset with at least 2–3 classes of text

Step 2: Data Preparation

Load dataset and inspect

- Clean text (remove stopwords, punctuation, lowercase, tokenization, lemmatization)
- Split into train/test sets (80/20)

Step 3: Feature Extraction

Convert text into numerical representation using at least two methods:

- TF-IDF Vectorizer
- Word Embeddings (Word2Vec, GloVe, or pre-trained embeddings)

Step 4: Model Training

Train and compare at least two classifiers on both feature sets:

- Logistic Regression
- Random Forest / Naïve Bayes
- (Optional: LSTM/GRU if they want deep learning exposure)

Step 5: Model Evaluation and Comparison

- ➢ Evaluate both models using:
- ➢ Accuracy, Precision, Recall, F1-score
- ➢ Confusion Matrix

Step 6: (Optional) Web App Deployment Using Streamlit

- Build a simple Streamlit app where users can enter text and get predictions (Spam/Ham, Sentiment, etc.)

Deadline:
Submit within **7 days**

Tools to Use:

- Python
- NLTK / SpaCy / re (for text preprocessing)
- scikit-learn
- TensorFlow / PyTorch (optional for embeddings/LSTM)
- Streamlit (optional deployment)