**Data Mining Course – Project Assignment**

**Project Title:** **Author Classification Using Text Mining and Machine Learning Techniques**

**Objective:**

The goal of this project is to classify documents based on their authorship using different feature extraction techniques and machine learning models. Given a dataset containing documents from multiple authors, students will analyze and compare various methods to achieve the best classification performance.

**Tasks:**

Students are required to perform the following steps:

1. **Text Representation Methods:**
   * Apply **TF-IDF (Term Frequency-Inverse Document Frequency)** to extract word-based features. Each word should be considered as a term.
   * Instead of using just words this time utilize **word-based n-grams**, specifically **2-grams and 3-grams**, as separate feature sets.
   * Utilize **character-based n-grams**, specifically **2-grams and 3-grams**, as separate feature sets.
   * Implement **BERT deep learning model** for text representation.
2. **Classification Algorithms:**
   * Train and evaluate the following classification algorithms on the extracted features:
     + **Random Forest (RF)**
     + **Support Vector Machine (SVM)**
     + **XGBoost (XGBOOST)**
     + **Naïve Bayes (NB)**
     + **Multi-Layer Perceptron (MLP)**
     + **Decision Tree (DT)**
3. **Evaluation Methods:**
   * The dataset should be evaluated under this experimental settings:
     + **80%-20% train-test split**
   * The classification performance should be measured using the following metrics:
     + **Accuracy**
     + **Precision (P)**
     + **Recall (R)**
     + **F1-score (F)**

**Deliverables:**

* A well-structured project report including:
  + Introduction to the problem and its importance.
  + Description of the dataset and preprocessing steps.
  + Explanation of the feature extraction techniques (TF-IDF, n-grams, BERT).
  + Implementation details of the machine learning models.
  + Experimental setup and evaluation methodology.
  + Performance comparison of the models with detailed results.
  + Conclusion summarizing key findings and insights.

**Submission Requirements:**

* The project report should be submitted as a **WORD document** with clear explanations and properly formatted tables and figures.
* Source code and dataset preprocessing scripts should be included in a **ZIP file** or a **GitHub repository link**.
* Ensure that all results are reproducible and that the methodology is clearly explained.

This project aims to provide hands-on experience in **text mining, machine learning, and deep learning-based text classification**, allowing students to explore real-world challenges in author attribution and automated text classification.

You may carry out the project as a group of up to **five** people. If you prefer, you may also do it individually.

Submission Date: May 7 , 2025

Good Luck.

Assoc.Prof.Dr.Aysun GÜRAN