Artificial Intelligence

Semester-end project assignment

Perform sentiment classification on the dataset provided with this document (*dataset_elec_4000.csv*), using two different machine learning algorithms, and compare their performance.

1. Students must submit their well-documented source code.
   - The dataset involves text reviews with ratings corresponding to the sentiment of reviews. Rating "1" corresponds to positive, and rating "0" corresponds to negative sentiments. For example, Review: *Doesn't fit my model as advertised. Ordered over holiday didn't know till i got home and opened it when it was too late to return. Rating: 0*
   - Although the recommended programming language for this assignment is Python, students may use any language of their choice.
   - Students are not required to make the algorithms from scratch, and may use any libraries and reuse any code found on the internet, textbooks, etc.
   - Students must explain all functions in their own words using comments (using functions from libraries) or docstrings (user-defined functions).
   - At least one of the algorithms must be a traditional machine learning method (e.g. Naive Bayes, Support Vector Machines, Random Forest, etc.). Although not required, students may use deep learning for the second algorithm (extra points).
   - The program(s) output must include the following: Name of the algorithm(s), accuracy, precision, recall, F1 score.
   - The program(s) must work with the original file (including file name).
   - The main file(s) must be named: <*studentname*>_<*algorithm(s)*>.<*extension*>, for example, KovacsMate_randomforest.py, KovacsMate_supporvectormachines_randomforest.py

2. Students must submit a report explaining the project and summarizing the results.
   - The report must be written in English and should have a logical structure (with sections). For example, *Introduction, Methodology, Experiments, Results, Conclusions*. Using tables, diagrams, and other figures for illustration purposes is encouraged (but all must be original).
   - The report must explain the domain (sentiment analysis), the dataset and the classification problem, potential preprocessing and feature engineering steps, and compare the acquired performance metrics of the two algorithms.
   - The length required is 3000 words, with a maximum tolerance of +10%. Only the main text counts (e.g. student name, references, etc. do not count).
   - The document must include student name, student ID, course and assignment name.
   - The only accepted format is PDF, and the report must be named: <*studentname*>*_finalreport.pdf*, for example, KovacsMate_finalreport.pdf