Fall 2023: CSCI 4/5587 Programming Assignment #3

DUE: Saturday, Nov 25, 2023 (**Softcopy** @ 11 PM)

Instructions

- **All work must be your own other than the instructor-provided data/code and hints to be used. You are not to work in teams on this assignment.**

Description

Training Dataset: A movie review dataset has been collected for sentiment analysis (see http://www.cs.cornell.edu/people/pabo/movie-review-data/). The dataset has been grouped into positive and negative classes (check Canvas for the dataset).

Task [Marks 100]: Develop the **analysis report** as described below and submit it: As demonstrated and discussed, the development of NASA's patent classifier for the fifteen-class classification problem, here similarly for the assignments, we will need to do the following steps and develop the **analysis report**:

i) [**10** points] Given the dataset, use Weka's TextDirectoryLoader to build up the initial ARFF file, containing the movie-review-text as strings and the output class {positive, negative}. Add the initial class distribution in the report. Also, report the required average conversion time in this step, describing how you obtained the average conversion time.

ii) [**15** points] Convert the text string to the most useful vector using Weka's unsupervised filtering tool: StringToWordVector. Report the chosen parameters, explain their roles, and justify your selections. Also, report how many words you have collected in this step.

iii) [**15** points] Using Weka's supervised filter, apply '**infoGainAttributeEval**' with '**Ranker'** having a threshold value set to 0.0. Report the total words retained for classification and write the first 10 words with their information-gain values.

iv) [**20** points] Run 10 different classifiers and measure their performances using 10 Fold Cross-Validation (FCV). Report all their performances (accuracy in %), including the confusion matrices. You must include the Naïve-Bayes approach as one of the 10 classifiers.

v) [**20** points] Report the best method with its parameter(s) you have found, including the performance-evaluation matrices. Explain why you think your selected top method is the best method out of the 10 methods you tried.

vi) [**20** points] Review literature to explain '**infoGainAttributeEval**' in detail and cite the relevant reference(s). Submit the copies of the paper(s) you have cited to explain '**infoGainAttributeEval**.'

--- x ---