ADVANCED MACHINE LEARNING

END TERM EXAMINATION

Start Time: 06-02-2023 04:00 PM
End Time: 07-02-2023 06:00 PM **Max Marks:** 150 marks

Note: The duration of the exam is 24 hours. An extra 2 hours have been given for submission. Submission beyond the end time would lead to the cancellation of the assignment.

PROBLEM: Developing an SMS Spam Filter

The SMS Spam Collection v.1 (hereafter the corpus) is a set of SMS-tagged messages that have been collected for SMS Spam research. It contains one set of SMS messages in English of 5,574 messages, tagged according being ham (legitimate) or spam.

You are required to perform the following (not limited to) activities:

1. **Reading the Data**
   Read the data in a Data Frame with two columns named SMS and Labels.

2. **Text Pre-processing** **[10 marks]**
   Use appropriate text pre-processing techniques to clean the texts.
   Be creative with your approaches here; each step needs to be justified.

3. **Data Exploration** **[10 marks]**
   Explore the word frequencies of the SMS in each category. Use word clouds for this. Explain your observations.

4. **Vectorization** **[40 marks]**
   Each SMS given in the corpus must be expressed in a vectorized form. Here are the four ways of doing so.
   a) Vectorization using TF.
   b) Vectorization using TF-IDF.
   c) Use doc-to-vec model using Google's word-to-vec model. (you may use any other embeddings as well like ELMO, etc.)
   d) Generate some heuristic features like "presence of phone number', etc.

5. **Dimension Reduction** **[40 marks]**
   a) Use PCA for dimension reduction. Keep the number of components that explains at least 95% of the overall variance.
   b) Visualize the data across the first two PCs. Use colour encodings for the labels for better visualization. (You may play across other PCs as well)
   c) You may try other heuristic approaches for dimension reduction as well like dropping texts that appear in less than p% of the documents. (This would bring you some brownie points)

6. **Spam Filters** **[50 marks]**
   a) Use an appropriate model-building framework to experiment with machine learning models like Logistic Regression, Decision Tree, Bagging, Boosting, Random Forest, SVM, etc. on the various vectorized representations you have created in stage 4.
   b) Use appropriate hyperparameter tuning techniques to control underfitting and overfitting.
   c) Report the confusion matrix, accuracy, precision, recall and F1-score of spam detection for each experiment you conduct.

SUBMISSIONS

1. **Google Colab File**
   Keep the following things in mind while creating your Google Colab file:
   a) The notebooks must be divided into sections & subsections.
   b) Every section and subsection must have a detailed introduction and methodology.
   c) All coding kernels must have detailed comments.
   d) The notebook must end with a proper conclusion.

2. **Slide Decks**
   - Slide 1: Personal Details
   - Slide 2: Data Pre-processing flow diagram
   - Slide 3: Word Clouds (Detail your observations from the word clouds)
   - Slide 4-7: Pipelines (or flow diagrams) for all four vectorization approaches.
   - Slide 5-8: Dimension Reduction & Visualization (for all 4 data frames)
   - Slide 9 & Onwards: A list of competing models (at least 10) with a briefing on the aim, experiment, observations and conclusions. (Note – one model on one slide)
   - Last slide: conclusion.

File nomenclature (for all files): Roll1-Roll2-Roll3-AML-ET

Note that Roll1 should be the lowest of all the three roll numbers, followed by roll2 and roll 3.

3. **Video Presentation** (To be submitted on YouTube)
   A video presentation of not more than 10 mins should be created and submitted on YouTube. The ***submission deadline for this work is 09-02-2023 11:59 PM.*** Note that each team members must present.