DSAI4205 - Big Data Analytics
Instructor: Dr. Ken FONG

Group Project Assignment (25%)
This project involves a **system development** task. Each team will develop a sentimental analysis on text data. The task details are as follow:

Project Overview:
Your group will be given a business textual dataset, with a standardized data splitting and evaluation metric. Your task is to develop a "good-performing" decision support system that classifies text.

1. Build a **baseline** model (i.e., a model with basic setup and no fine-tuning, e.g., logistic regression)
2. Compare with **at least N other data analytics models** (e.g., Gradient boosting tree, decision tree, etc) and **word embeddings models** (BERT, GloVE, Word2vec, etc)
3. Develop methods to improve your baseline models (e.g., ensemble models, fine-tune different hyper-parameter for the models above, etc) and **report the best score**.
4. **Visualize with at least N data statistics**. There are many ways that can describe patterns found in data, including chart and figures (e.g., bar charts, line graphs, word cloud, word distribution, etc), as well as statistical metrics like central tendency (mean, mode and median) and dispersion (range, variance, maximum, minimum, quartiles), correlation and standard deviation.

(*N* correspond to your number of group members)

Note that higher marks will be given for developing novel and viable methods that are not taught in class

In your code submission part (i.e., Jupyter Notebook file), they need to cover **the code implementation for Point 1-4**; and **a short table and/or text (less than 300 words)** that describe Point 1-4 (e.g., what model, hyper-parameter, visualization techniques you have tried and/or their scores and details). You can elaborate more detail in your report part. And highlight the main point in your power point (refers to the Submission Section below).

You can use any visualization and model improvement techniques. However, you have to use **the SAME data splitting and evaluation metric throughout the project** (marks will be deduced if these two conditions are found changed in the submission).

Each student needs to upload **a work distribution file** like below in word file to the Blackboard to indicate the work distribution of you and your teammate (you don't need to let your group mate to know the work distribution you feel)

Team X:

Members (Full name)	Student ID	Work Distribution
Chan Tai Man	000000	20%
Wong Tai Man	000001	20%
Chen Tai Man	000002	20%
Tsang Tai Man	000003	20%
Ng Tai Man	000004	20%

Marking Scheme for Group project (20%)

- Implementation of a big analytical data management system and decision support system using professional software (10%): Whether your group have a good-performing system and a comprehensive evaluation
- Problem definition (demonstrate the understanding of the problem and formulate alternative solutions) (5%): Within the scope of the Big-Data area, whether your group have clearly describe the problem, list out related constraints, and able to formulate alternative solutions.
- Creative solution design (5%): Whether you can link your Big-Data solution back to the context of the company case

Marking Scheme for Presentation (5%)

- Appropriate time allocation and pace (0.5%): Did your group allocate time appropriately, and mange time effectively, with smooth progression? Did your group use appropriate pace? Did the presentation start punctually?
- Clear, logically organized and relevant content (2%): Was information included always relevant? Were presented points clearly stated and developed? Did the materials flow extremely well? Were the materials well organized? Were there any ambiguities are left unexplained?
- Uses good body language, eye contact, and appropriate voice tone (1%): Did the group show balanced posture, enthusiasm and confidence? Did the group make good eye contact with audience? Did the group use voice tone effectively?
- Gains/holds attention (1%): Did the group provide good motivation to engage the audience's interest? Did the group present the contents in a manner that captivates the audience's attention?
- Clarity of speech/Accuracy of grammar & pronunciation (0.5%): Was the voice consistently comprehensible? Were grammar and pronunciation accurate?

Submission (23 Nov 2025 23:55 HK Time)

Each **team** should find a representative to submit three files onto the blackboard:

1. A Jupyter Notebook file containing all deliverables regarding the system development task (e.g., the model performance, evaluation description and visualization)
2. A written report in MS Word file (**2 pages max**) that provides a background description of your system, how you approach the problem, and a critical evaluation of the system.
3. A PowerPoint file that you use in the presentation covering highlights in (1) and (2).