MCSD2123 – MASSIVE DATA MINING AND STREAMING
GROUP ASSIGNMENT
(Must complete Part A and B)

PART A: CLASSIFICATION USING TEXT

INSTRUCTIONS:

1. Each group is required to identify **ONE (1)** case that relate to Supervised Machine Learning with application on Textual Data. (Example: sentiment analysis, text classification, etc) in any domain.
2. Then identify suitable dataset(s) to be used for the chosen case (you may use more than **ONE (1)** dataset if needed).
3. Study and perform Exploratory Data Analysis (EDA) to understand the datasets.
4. Explore and select at least **TWO (2)** supervised machine learning algorithms for training the models using the identified dataset(s). Make sure datasets are properly prepared and processed.
5. You should perform appropriate evaluation and comparison of the performance of all models.
6. Conclude your findings. In your selected case and data, which model performs the best? Provide your discussion.

DELIVERABLES:

1. Presentation slides that with your findings. The slides should cover the following:
    a. Executive summary
    b. Introduction / Background
    c. Objectives
    d. Dataset and EDA
    e. Methodology
        i. Overview of solution
        ii. Preprocessing
        iii. Model Training
        iv. Model Evaluation
    f. Results and Findings
    g. Conclusions

    ** *You may use more than one slide for explanation of any of the mentioned topics above.*

2. implementation in Python Notebook (.ipynb)

3. Document that states the distribution of work among group members. (In PDF)

PART 2: CLUSTERING OF TEXT

Clustering of text is an approach that can be used to explore similarities between different sentences, identify common words that grouped into similar clusters without any prior context in categorizing the text.

INSTRUCTIONS:

1. In this case, using the same chosen dataset above, explore and perform clustering using **TWO (2)** different clustering algorithms. Then analyze the resulting clusters and interpret your findings.
2. You should design your experiment accordingly, then evaluate and compare the performance of the clusters from the chosen algorithms. Which clustering result is the best in your case?
3. Analyze and discuss the clusters from the best clustering results. You may use any interesting visualizations in supporting your analyses and discussion.

DELIVERABLES:

1. Presentation slides that with your findings. The slides should cover the following:
    a. Executive summary
    b. Introduction / Background
    c. Objectives
    d. Methodology
    e. Results and Findings
    f. Conclusions

2. Implementation in Python Notebook (.ipynb)

ASSESSMENT RUBRIC

CRITERIA	POOR (0-3)	MODERATE (4-6)	GOOD (7-10)	WEIGHT* (%)
Requirement	Submission incomplete.	Submission requirements are partially fulfilled, some files missing.	All requirements fulfilled.	5.0
Executive Summary	Executive Summary is incomplete/irrelevant.	Executive Summary is limited.	Executive Summary is well structured and clearly explained.	5.0
Data Preparation	Data preparation is not explained.	Minimal explanation of data preparation, some information is unclear	Complete explanation of data preparation.	15.0
Methodology	Methodology is not explained. No justification of the choice of thresholds.	Methodology is not clearly explained. Some of the explanation of choice of threshold and performance measure is unclear.	Methodology is clearly explained with proper justification of the choice of threshold in the study and the performance measure.	20.0
Results	Results are not shown and explained.	Results are shown with minimal explanation. Some results are incorrect. Performance justified in limited aspects.	All results are correctly shown with explanation. Performance justified in different aspects.	25.0
Interpretation	No interpretation is provided. No explanation is provided.	Interpretation provided but not clearly explained.	Clear interpretation and explanation provided.	25.0
Conclusion	No conclusion provided	Some explanation in conclusion section, but no conclusive statement based on findings.	Conclusion is clearly written according to findings.	5.0
				100.0