MOD006562 (Machine Learning)

Assessment Element 010

This module has one assessment element which carries 100% of your total marks and reflects the entire **learning outcomes** of this module as below:

No.	Type	On successful completion of this module the student will be expected to be able to:
1	Knowledge and understanding	Demonstrate an understanding of the key concepts underpinning machine learning
2	Knowledge and understanding	Demonstrate knowledge of most common machine learning algorithms
3	Intellectual, practical, affective and transferrable skills	Identify the most appropriate algorithm for the problem to be solved
4	Intellectual, practical, affective and transferrable skills	Train and validate a range of models for supervised learning

Objectives

This assignment requires you to implement and evaluate multiple machine learning models on a given dataset.

Assessment Description:

This assessment requires you to

(i) Implement and evaluate various machine learning models and vectorisation techniques on a given dataset
(ii) Write a reflective report describing your experiments and critical analysis (90%)
(iii) Present your work orally aided by PowerPoint slides in a short presentation session n (10%)

1. Implementation

This assessment aims to evaluate your practical skills in implementing an end-to-end machine learning pipeline based on the provided dataset of News Categories (News_Categories.csv). This dataset consists of more than 200,000 news headlines that categorized based on their topics into 41 different categories as below:

'CRIME', 'ENTERTAINMENT', 'WORLD NEWS', 'IMPACT', 'POLITICS', 'WEIRD NEWS', 'BLACK VOICES', 'WOMEN', 'COMEDY', 'QUEER VOICES', 'SPORTS', 'BUSINESS', 'TRAVEL', 'MEDIA', 'TECH', 'RELIGION', 'SCIENCE', 'LATINO VOICES', 'EDUCATION', 'COLLEGE', 'PARENTS', 'ARTS & CULTURE', 'STYLE', 'GREEN', 'TASTE', 'HEALTHY LIVING', 'THE WORLDPOST', 'GOOD NEWS', 'WORLDPOST', 'FIFTY', 'ARTS',

MOD006562 (Machine Learning)

Assessment Element 010

'WELLNESS', 'PARENTING', 'HOME & LIVING', 'STYLE & BEAUTY', 'DIVORCE', 'WEDDINGS', 'FOOD & DRINK', 'MONEY', 'ENVIRONMENT', 'CULTURE & ARTS'

This assessment requires you to:

- Import and warehouse the given dataset into your workspace using efficient and suitable data structures.
- Conduct comprehensive pre-processing process that may involve operations such as data cleaning, reshaping, resizing, dealing with missing values and other commonly practised industry-standard data wrangling operations.
- Besides general-purpose data wrangling operations (above), you must conduct comprehensive NLP specific data preparation operations such as cleaning, case normalisation, stop-word removal, lemmatising and stemming, etc.
- Prepare your training, validation and testing sets and their corresponding ground truth labels.
- Perform comprehensive statistical analysis of the given dataset. This may include simple statistical analysis, histogram analysis, outlier analysis, correlation analysis, etc.
- Develop at least 3 different natural language vectorization (feature extraction) techniques.
- Develop a minimum of 3 different supervised machine learning classification models and train, validate, and test these models using feature vectors produced as a result of vectorization operations.
- Express your results in both quantitative and qualitative manners. Use commonly practised industry-standard evaluation metrics including, accuracy, F-1 score, Precision, Recall, and confusion matrix to express your results.

Remarks:

- Python version 3.6 or above must be used in this assessment.
- You are free to choose your desired IDE (However, preference is Jupyter Notebook)
- Your code must be fully commented. This demonstrates your understanding of the proposed work.
- The novelty and originality of your solution are extremely important.

2. Reflective Report

You should pair your implementation with a reflective report of no more than 2500 words that describes your proposed solution's components, used algorithms, processes, parameters and outlines their rationale and justification. For example, you should explain what classification models you chose for this task and why? what training parameters are paired with those models and why? You should provide such a narrative for every single component

of your proposed solution. You should apply the knowledge you gained from the lectures and practical sessions and complement it with your own research in order to demonstrate your understanding of the subject material. You are expected to thoroughly and accurately reference your report using the Harvard referencing style expected by Anglia University. The report should contain a minimum of 10 references from authoritative sources.

Details on Harvard referencing can be found on the library website:

https://libweb.anglia.ac.uk/referencing/harvard.htm

Please use the provided report template as the baseline.

3. Presentation

A short presentation (8 min presentation + 2 min Q&A) session will be scheduled toward the end of this trimester (Week 10, 11, 12). You must briefly talk about your proposed solution, its components and their rationale and demonstrate your data analysis results and findings. Provided feedback can help you refine your work and address its shortcomings.

MOD006562 (Machine Learning)

Assessment Element 010

Guidelines:

- Submission Deadline: Please refer to the Canvas Assessment page.
- Assessment Type: **Individual** – Coursework
- Report length: The expected report word count is 2500 words (equivalent), however 15% leeway, either way, will be allowed. Please include your word count on the cover page.
- Report Format and Naming Convention: StudentID.PDF or DOCX
- Code File Format: Ideally Jupyter Notebook/PDF or zip file
- Submission Platform: Canvas
- Important Note: The report language must be formal, written in the third person. have all figures and tables correctly labelled and referenced and be presented in a structured and meaningful way, with consideration for grammar, punctuation, and spelling.
- **Important Note:** Ensure your student ID is listed on ALL pages' header. Submission should be anonymous, so DO NOT include your name on any pages of your submission.
- **Important Note:** All figures and diagrams provided in the report must be in high resolution, clear, and readable.
- **Important Note:** References and in-text citations (if any) must be presented using the Harvard/ARU referencing style. Please refer to the following for more info:

  https://library.aru.ac.uk/referencing/harvard.htm

- **Important Note:** The report must be submitted within the university agreed timescales. Please refer to the followings for late submissions T&C:

  https://web.anglia.ac.uk/anet/academic/assess/latesubmission.phtml

MOD006562 (Machine Learning)

Assessment Element 010

Marking Criteria:

Report and Code: *To produce quality work in general, (i) give attention to report structure to make all sections readable and coherent in flow, (ii) use relevant academic references to support your content and (iii) use good coding practices.*	
Data pre-processing, preparation, and wrangling *Marks will be given for the depth of data exploration, pre-processing and your reflection on it. Your analysis will be regarded higher if supported with a good commentary/reflection/explanation.*	15
Statistical analysis *Marks will be given for statistical analysis and visual exploration via plots. Your plots and statistics will be regarded higher if supported with a good commentary/reflection/explanation.*	10
Training, validation, and testing set preparation *This section requires you to provide code snippets for train and test splits and explaining parameter choices.*	05
Feature extraction (vectorization) *This section requires you to provide details of your vectorization techniques. Explain if you performed any methods for dimensionality reduction etc. Marks will be given to your code's execution and functionality, design and efficiency, accuracy and performance, novelty and originality and your report contents and structure.*	10
Classification (training, validation, and testing) *Provide details of the chosen models and their implications. Marks will be given to your code's execution and functionality, design and efficiency, accuracy and performance, novelty and originality and your report contents and structure.*	30
Qualitative and quantitative results' representation *Critically discuss the results and any factors affecting the results. Marks will be given to your code's execution and functionality, design and efficiency,*	20

MOD006562 (Machine Learning)

Assessment Element 010

accuracy and performance, novelty and originality and your report contents and structure.	
Oral Presentation	
Presentation: - 10 minutes long (8 + 2 minutes for questions) - PowerPoint based Presentation - Explain your dataset analysis - I appreciate that presentations will be conducted from Week 10 and you may not have completed your tasks. You can explain whatever has been done so far and what the plan for rest of the project is. - Presentation feedback will help you to understand if you are on the right track and your work is going to meet the requirements. - The marks will be given for o Visual and logical structure of the slides (2) o Clarity of the content (2) o Pace (1) o Explanation (2) o Question Handling (2) o Time Management (1)	10

How to Submit:

Please submit Report and Code separately as two files.