	PROJECT 3	Mark Yung
1	**Problem Statement**	
	Is it clear what the student plans to do?	3.0
	What type of model will be developed?	3.0
	How will success be evaluated?	0.0 [1]
	Is the scope of the project appropriate?	3.0
	Is it clear who cares about this or why this is important to investigate?	3.0
	Does the student consider the audience and the primary and secondary stakeholders?	3.0
	Problem Statement - Averaged Marks (out of 3)	3.0
2	**Data Collection**	
	Was enough data gathered to generate a significant result?	3.0
	Was data collected that was useful and relevant to the project?	3.0
	Was data collection and storage optimized through custom functions, pipelines, and/or automation?	3.0
	Was thought given to the server receiving the requests such as considering number of requests per second?	3.0
	Data Cleaning and EDA - Averaged Marks (out of 3)	3.0
3	**Data Cleaning and EDA**	
	Are missing values imputed appropriately?	3.0
	Are distributions examined and described?	3.0
	Are outliers identified and addressed?	1.0 [2]
	Are appropriate summary statistics provided?	3.0
	Are steps taken during data cleaning and EDA framed appropriately?	3.0
	Does the student address whether or not they are likely to be able to answer their problem statement with the provided data given what they've discovered during EDA?	3.0
	Data Cleaning and EDA - Averaged Marks (out of 3)	2.7
4	**Preprocessing and Modeling**	
	Is text data successfully converted to a matrix representation?	3.0
	Are methods such as stop words, stemming, and lemmatization explored?	2.0
	Does the student properly split and/or sample the data for validation/training purposes?	1.0 [3]
	Does the student test and evaluate a variety of models to identify a production algorithm (**2 Models should be used**)?	3.0
	Does the student defend their choice of production model relevant to the data at hand and the problem?	1.0 [4]
	Does the student explain how the model works and evaluate its performance successes/downfalls?	0.0
	Preprocessing and Modeling - Averaged Marks (out of 3)	1.7
5	**Evaluation and Conceptual Understanding**	
	Does the student accurately identify and explain the baseline score?	0.0
	Does the student select and use metrics relevant to the problem objective?	1.5
	Does the student interpret the results of their model for purposes of inference?	1.0
	Is domain knowledge demonstrated when interpreting results?	3.0
	Does the student provide appropriate interpretation with regards to descriptive and inferential statistics?	3.0
	Evaluation and Conceptual Understanding - Averaged Marks (out of 3)	1.7
6	**Conclusion and Recommendations**	
	Does the student provide appropriate context to connect individual steps back to the overall project?	3.0
	Is it clear how the final recommendations were reached?	3.0
	Are the conclusions/recommendations clearly stated?	1.0
	Does the conclusion answer the original problem statement?	3.0
	Does the student address how findings of this research can be applied for the benefit of stakeholders?	1.0
	Are future steps to move the project forward identified?	3.0
	Conclusion and Recommendations - Averaged Marks (out of 3)	2.3

7	**Project Organization**	
	Are modules imported correctly (using appropriate aliases)?	2.0 [7]
	Are data imported/saved using relative paths?	3.0
	Does the README provide a good executive summary of the project?	3.0
	Is markdown formatting used appropriately to structure notebooks?	3.0
	Are there an appropriate amount of comments to support the code?	3.0
	Are files & directories organized correctly?	2.5 [8]
	Are there unnecessary files included?	3.0
	Do files and directories have well-structured, appropriate, consistent names?	3.0
	Project Organization - Averaged Marks (out of 3)	2.8
8	**Visualizations**	
	Are sufficient visualizations provided?	3.0
	Do plots accurately demonstrate valid relationships?	3.0
	Are plots labeled properly?	3.0
	Are plots interpreted appropriately?	3.0
	Are plots formatted and scaled appropriately for inclusion in a notebook-based technical report?	3.0
	Visualizations - Averaged Marks (out of 3)	3.0
9	**Python Syntax and Control Flow**	
	Is care taken to write human readable code?	3.0
	Is the code syntactically correct (no runtime errors)?	3.0
	Does the code generate desired results (logically correct)?	1.0 [9]
	Does the code follow general best practices and style guidelines?	2.0 [10]
	Are Pandas functions used appropriately?	3.0
	Are \`sklearn\` and \`NLTK\` methods used appropriately?	3.0
	Python Syntax and Control Flow - Averaged Marks (out of 3)	2.5
10	**Presentation**	
	Is the problem statement clearly presented?	2.5
	Does a strong narrative run through the presentation building toward a final conclusion?	2.5
	Are the conclusions/recommendations clearly stated?	2.5
	Is the level of technicality appropriate for the intended audience?	2.5
	Is the student substantially over or under time?	2.5
	Does the student appropriately pace their presentation?	2.5
	Does the student deliver their message with clarity and volume?	2.5
	Are appropriate visualizations generated for the intended audience?	2.5
	Are visualizations necessary and useful for supporting conclusions/explaining findings?	2.5
	Presentation - Averaged Marks (out of 3)	2.5
Total:	**Overall Marks (out of 30)**	**25.2**
	Average Marks (out of 3)	**2.52**
30	**Percentage**	**84.0%**

Comments	[1] No indication which metrics will be used to evaluate models [2] It would be better, and important to analyse and drop duplicate/spam documents (row instead of word token) where they are not needed. This can give more meaningful EDA insights (e.g. no top few words from spam), affects vectorisation (i.e. TF-IDF that look at corpus), and affects prediction [3] Train test split correctly, however unable to find target variable value being encoded before modelling. [4] Only mentioned model is selected due to best model for distinguishing between subreddits. Could elaborate your analysis/evaluation of models according to chosen metrics (unknown), in combination with other consideration, e.g. stakeholder consideration (interpretability vs prediction power), computation time [5] Did not indicate what metrics/consideration will be used for model evaluation in relation to problem statement/business case - assumed AUC ROC from brief desscription [6] Final model is logistic regression that offers easy interpretability - could describe your interpretation on model fit, accuracy, generalisation, extract key features using features coeff [7] Import libraries needed for respective notebook (e.g. see Notebook 1) [8] Recommended to place multiple notebooks into a folder [9] Warning message for modelling: max_df result in less than min_df. This means invalid parameters to gridsearch - see notebook 3, above “3.3 Model Result” for explanation [10] Check target variable distribution for classification prediction to verify if you're working with balanced/imbalanced data - serve as a consideration if Accuracy metric is favourable or otherwise. Misc - Notebook is lacking some details available in README, i.e. MLflow result, conclusion section, etc. Please include into notebook for complete narrative flow. You may consider drafting all details within notebook, then create README content by extracting and distillling from notebook