NLP Project Preference Document				**IT9002**
Mentor : Sini Raj Pulari				
Project Name:				
Student Name:	**Manal Talal Ameen**			**Student I ID: 12011218**
			NLP Project Log File	
Date	**Week No**	**Task Name**	**Work done During the week (Bullet points /Description)**	**Issues Experienced if any**
18/10/2025	**4**	**Exploring topics recommended**	· Went through the list of recommended topics of NLP projects that are stated by the tutor. · Strong themes which are applicable in my marketing experience and personal interest. · Chose to research the area of Sentiment Analysis and Natural Language Processing in marketing. · Scanned potential data sets and selected Amazon Product Reviews as the most valuable data set: https://nijianmo.github.io/amazon/index.html"	· Several URLs from the list were inactive or closed immediately when accessed. · Difficulty locating stable and accessible datasets at first.
23/10/2025	**5**	**Practiced installing toolkit and NLP Lab**	· Practiced installing and importing NLP libraries (NLTK, Sklearn, Pandas). · Set up Google Colab environment connected to Google Drive. · Repeated class lab exercises (tokenization, stemming, lemmatization, vectorization). · Tested sample code to build confidence in preprocessing steps.	· Faced initial errors in mounting Google Drive and missing library installations. · Required time to understand file paths and toolkit commands.
30/10/2025	**6**	**Generic research**	· Conducted general research on sentiment analysis in digital marketing. · Studied how NLP is used to analyze consumer behavior and product reviews. · Explored research papers on AI-generated content and Gen Z purchase intention for foundational understanding.	· Overwhelmed by the amount of research available; needed to narrow focus.
13/11/2025	**7**	**decided on a topic**	· Finalized project topic: Sentiment Analysis on Amazon Product Reviews using NLP. · Matched topic to project requirements (classification, preprocessing, vectorization). · Started outlining the project flow based on Tasks 1–6.	· Uncertainty about whether to relate it directly to my research-methodology topic; decided to follow the tutor's recommended topic for alignment.
20/11/2025	**9**	**collected dataset from kaggle**	· Downloaded structured Amazon Reviews dataset from Kaggle. · Imported the dataset into Google Colab for inspection. · Checked column names, missing values, and dataset size. · Verified dataset suitability for sentiment classification.	· Faced encoding errors and needed to load dataset using 'ISO-8859-1' encoding. · Required cleaning due to mixed formats and text noise.
27/11/2025	**10**	**Exploratory Data Analysis (EDA**	· Performed EDA on the Amazon dataset using pandas and Matplotlib. · Generated visualizations for sentiment distribution (bar chart + pie chart). · Analysed review length distribution and word frequency. · Identified patterns in positive vs negative reviews. · Captured screenshot evidence for the report (Task 2).	· Some visualizations had formatting issues; required multiple attempts. · Review lengths varied widely, making interpretation slightly difficult.
2/12/2025	**11**	**Text processing**	· Removed URLs, special characters, numbers and additional spaces in a raw review text. · Fixed all text to lower case. · NLTK tokenized review to individual words. · Noise reduction by applied stopword removal. · WordNet Lemmatizer and implemented stemming (Porter Stemmer). · Comparisons made with tokenized, stemmed and lemmatized outputs. · Prepared documentation in form of created tables and screenshots.	· Initial errors due to missing NLTK resources. · Required multiple downloads and environment resets in Google Colab.

10/12/2025	**12**	**Feature extracting**	· TF -IDF vectorization was used to convert cleaned textual data into numerical features. · The feature set was restricted to ensure that dimensionality and computational cost was controlled. · The shape and sparsity of the resulting matrix of features were verified. · These features were ready to be used in machine-learning models. · The entire procedure of feature- extraction was reported.	· Further experimentation was needed to adjust TF-IDF parameters. · The first high dimensionality increased memory expenses.	
23/12/2025	**14**	**Sentiment label creation**	· Transformed numeric star ratings into sentiment scores: Negative (1 -2), Neutral (3), Positive (4 -5) · Introduced an individual sentiment labeling feature. · Confirmed the frequency counts of the classes. · Plotted the sentiment distribution with the bar charts. · Added screen shots of labeled data in the report.	· The imbalance of classes is extreme particularly in neutral reviews. · Special caution must be exercised when subsequent evaluation is done.	
29/12/2025	**14**	**Model Training**	· Trained 2 classical machine-learning models: · Multinomial Naive Bayes · Logistic Regression · The two models were fed with TF-IDF features. · The data was divided into a training and testing sample. · I was able to come up with projections to assess performance. · I made a comparison of the performance of the two models at the baseline.	· The first problem was confusion due to overlapping variable names of models during prediction. · To address this, we require more categorical naming of variables to debug Colab.	
30/1/2025	**15**	**Advanced NLP Analysis & Report Writing**	· Big-gram frequency analysis using Countvectorizer. · The most common pairs of words were identified. · Part of speech tagging with NLTK. · Made sense of the linguistic patterns that were depicted by the customer reviews. · Completed the discussion, limitations and conclusion sections. · Went through the report and made sure that it is in line with the marking structure.	· The POS tagging output was long hence we needed to report only what was important. · A lot of caution had to be taken in balancing between explanations and screenshots.	
4/1/2026	**15**	**Code debugging and validation**	· Debugged Google Colab notebook to resolve tokenization and preprocessing errors. · Verified consistency between preprocessing steps and report explanations. · Re-ran key cells to regenerate correct outputs and figures. · Validated Logistic Regression results and confusion matrix outputs. · Cleaned unnecessary or duplicate cells from the notebook.	· Errors caused by variable overwriting and missing NLTK resources. also I struggled everytime when i reconnect to the colab notebook · Required careful tracing of preprocessing order.	
6/1/2016	**15**	**Finalizing**	· Finished discussion, limitations and conclusion parts of the report. · References were reviewed and proper academic citation formatting established. · Matched screen shots with matching code results. · Ready finished ZIP file with: · Project report (Word) · Google Colab notebook (.ipynb) · Data and supplements. · GitHub webpage Approval of project against submission checklist.		