Data Collection and Preprocessing Phase

Date	20 November 2024
Team ID	SWTID1727420425
Project Title	Analysis of amazon review using nlp technique
Maximum Marks	2 Marks

Data Quality Report Template

The Data Quality Report Template will summarize data quality issues from the selected source, including severity levels and resolution plans. It will aid in systematically identifying and rectifying data discrepancies. This template provides a comprehensive view of potential data quality issues in the Amazon reviews dataset and outlines technical solutions to address these challenges, ensuring better accuracy and consistency in the NLP analysis.

Data Source	Data Quality Issue	Severity	Resolution Plan
Amazon Reviews Dataset	Duplicate reviews (identical reviews posted multiple times).	Moderate	Deduplicate reviews based on a combination of review ID and text similarity (e.g., using a threshold for cosine similarity or exact matching).
Amazon Reviews Dataset	Missing or incomplete review data (e.g., empty reviews, missing ratings).	High	Implement a data cleaning step to filter out empty reviews and ensure that ratings are present. Use imputation techniques if needed for missing ratings

Amazon Reviews Dataset	Noisy or irrelevant reviews (e.g., reviews that don't provide useful sentiment or context, such as spam or promotional content).	High	Preprocess data to identify and remove non-relevant content using keyword filtering, text length checks, or external classifiers to detect spam.