**Data Collection and Preprocessing Phase**

|  |  |
| --- | --- |
| Date | 25 Jan 2025 |
| Team ID | 740678 |
| Project Title | Amazon Kindle Store Reviews Analysis |
| 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.

|  |  |  |  |
| --- | --- | --- | --- |
| **Data Source** | **Data Quality Issue** | **Severity** | **Resolution Plan** |
| Review Dataset 1 | Reviews may contain excessive emojis, HTML tags, or non-ASCII characters, leading to noise in sentiment analysis. | Medium | Use regex and libraries like BeautifulSoup and emoji to clean HTML tags and filter emojis/non-standard characters. Normalize the text using UTF-8 encoding. |
| Review Dataset 2 | Duplicate or near-duplicate reviews, often posted across multiple books or editions. | High | Apply text similarity techniques (e.g., cosine similarity, Jaccard index) and remove duplicates using fuzzy matching or MinHash-based deduplication. |
| Review Dataset 3 | Missing or null values in key fields like rating, review text, or review date. | High | Impute missing ratings with median or mean values; discard or flag entries with missing review texts. Validate date formats using datetime parsing. |
| Review Dataset 4 | Language inconsistencies (e.g., multilingual reviews not detected) affect NLP tasks. | Medium | Use language detection tools like langdetect or fastText to identify and segment reviews by language. Filter or separately analyze non-English content. |
| Review Dataset 5 | Fake or spam reviews (e.g., overly generic text, extreme ratings). | High | Use anomaly detection and classification models trained to flag spammy content based on patterns like review length, time patterns, and word repetitiveness. |