**Data Collection and Preprocessing Phase**

|  |  |
| --- | --- |
| Date | 18 June 2025 |
| Team ID | SWTID1750052396 |
| Sentiment Analysis of App Reviews | Analysis of medium app review from Google Play Store |
| 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** |
| dataset.csv | Missing values in reviewCreatedVersion | Moderate | Filled with mode using df['reviewCreatedVersion'].fillna(mode, inplace=True) |
| dataset.csv | Missing values in replyContent | Moderate | Filled with mode |
| dataset.csv | Missing values in repliedAt | Low | Filled with mode |
| dataset.csv | Missing values in appVersion | Low | Filled with mode |
| dataset.csv | Redundant columns like reviewId, at, repliedAt, replyContent, appVersion, reviewCreatedVersion | Low | Dropped them using df.drop(..., axis=1) |
| content | Inconsistent casing and special characters | Moderate | Lowercased all text and removed special characters using regex |
| content | Presence of stopwords | Low | Removed using NLTK’s stopword list |
| content | Unequal text lengths and sparsity | Moderate | Applied TfidfVectorizer with min\_df=2, ngram\_range=(1,3) |