**Data Collection and Preprocessing Phase**

|  |  |
| --- | --- |
| Date | 27 Sept 2025 |
| Team ID | LTVIP2025TMIDS63456 |
| Project Title | Analysis of medium app review from Google Play Store |
| Maximum Marks | 6 Marks |

**Preprocessing Template**

The images will be preprocessed by resizing, normalizing, augmenting, denoising, adjusting contrast, detecting edges, converting color space, cropping, batch normalizing, and whitening data. These steps will enhance data quality, promote model generalization, and improve convergence during neural network training, ensuring robust and efficient performance across various computer vision tasks.

|  |  |
| --- | --- |
| **Section** | **Description** |
| Data Overview | The dataset contains user reviews and associated sentiment labels (sentiment). Other unused metadata columns were dropped. |
| Loading Data | |  | | --- | |  |  |  | | --- | | Used pandas.read\_csv() to load the CSV file into a DataFrame. | |
| Handling Missing Data | Filled missing values in reviewCreatedVersion, replyContent, repliedAt, and appVersion using mode. Dropped unused columns like reviewId, at. |
| Text Cleaning | Texts were lowercased, punctuation removed using regex, and multiple spaces collapsed into one. |
| Stopword Removal | Eliminates common words (like "the", "is", "and") that don't add meaningful information, helping improve model focus and performance. |
| Vectorization | Converts text into numerical form (using TF-IDF) so machine learning models can process and learn from it. |
| **Data Preprocessing Code Screenshots** | |
| Loading Data |  |
| Handling Missing Data |  |
| Text Cleaning |  |
| Stopword Removal |  |
| Vectorization |  |