Data Collection and Preprocessing Phase

|  |  |
| --- | --- |
| Date | 15 October 2024 |
| Team ID | 739894 |
| Project Title | Toxic Comment Classification for Social Media using NLP |
| 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** |
| Comments Dataset | Imbalanced classes (toxic vs. non-toxic) | high | Use techniques like SMOTE (Synthetic Minority Over- sampling Technique) or re- sampling to balance classes. |
| Comments Dataset | Presence of null or missing comments | moderate | Remove rows with null entries or replace them using context-based data imputation. |
| Comments Dataset | Presence of offensive keywords in non-toxic class | high | Manual relabeling or use a semi- supervised learning approach to refine the labels. |

|  |  |  |  |
| --- | --- | --- | --- |
| Comments Dataset | High variance in comment length | low | Normalize lengths by truncating excessively long comments and padding short ones. |
| Comments Dataset | Mismatch in encoding or format | low | Convert all text data into UTF-8 encoding and standardize format during preprocessing. |