Data Collection and Preprocessing Phase

Date	26 November 2024
Team ID	SWTID1726490119
Project Title	Toxic Comment Classification for Social Media
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
Kaggle Dataset	Missing values in the comment_text column.	High	Remove rows with missing values or apply NLP preprocessing to handle empty comments (e.g., replace with "No text provided").
Kaggle Dataset	Imbalanced dataset across labels (e.g., toxic, severe_toxic, obscene, etc.).	High	Use techniques like oversampling (e.g., SMOTE), undersampling, or class weighting during training to handle imbalance.

Kaggle Dataset	Presence of special characters, URLs, and HTML tags in the comment_text column.	Moderate	Clean the text using preprocessing techniques such as text cleaning.
Kaggle Dataset	Categorical data in labels (e.g., binary values for toxic, severe_toxic).	Moderate	Ensure proper encoding is applied during model preprocessing. Convert labels to one-hot encoding if needed for multi-label classification.
Kaggle Dataset	Potential duplicates in the comment_text column.	Low	Check for duplicates and remove them to avoid redundant training data.
Kaggle Dataset	Mixed-case text in comment_text.	Low	Convert all text to lowercase for consistent NLP processing.