Data Collection and Preprocessing Phase

|  |  |
| --- | --- |
| Date | 3 october 2024 |
| Team ID | LTVIP2024TMID24818 |
| Project Title | Toxic Comments Classification For Social Media |
| Maximum Marks | 2 Marks |

# Data Collection Plan & Raw Data Sources Identification Report:

Elevate your data strategy with the Data Collection plan and the Raw Data Sources report, ensuring meticulous data curation and integrity for informed decision-making in every analysis and decision-making endeavor.

# Data Collection Plan:

|  |  |
| --- | --- |
| **Section** | **Description** |
| Project Overview | This project Classification of online toxic comments using machine learning algorithms is used to classify toxic comments. Toxic Comments Classification is an AI-powered web application that detects and categorizes harmful online comments. Utilizing Natural Language Processing (NLP) and Machine Learning (ML), The system classifies given data into 6 categories i.e. threat, insult, toxic, severe toxic, obscene, and identity hate. |
| Data Collection Plan | * Collect comments from social media platforms and online forums via APIs, web scraping, and manual annotation. * Ensure balanced dataset with 50:50 toxic and non-toxic comments, diverse in comment types (hate speech, harassment, spam). |
| Raw Data Sources Identified | Toxic Comments Classification on social media utilizes various raw data sources, including public datasets (Kaggle's Toxic Comment Classification Challenge, and social media platforms, social media moderation datasets, and hate speech databases, ensuring a comprehensive and diverse dataset for accurate toxic comment classification. |

**Raw Data Sources Report:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source Name** | **Description** | **Location/URL** | **Format** | **Size** | **Access Permissions** |
| Kaggle Dataset | You are provided with a large number of Wikipedia comments which have been labeled by human raters for toxic behavior. The types of toxicity are:   * toxic * severe\_toxic * obscene * threat * insult * identity\_hate | <https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge/data> | CSV | 55.18MB | Public |