Data Collection and Preprocessing Phase

|  |  |
| --- | --- |
| Date | 15 October 2024 |
| Team ID | 739916 |
| Project Title | Toxic Comment Classification for Social Media using NLP |
| Maximum Marks | 2 Marks |

**Data Collection Plan & Raw Data Sources Identification Template**

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 Template**

|  |  |
| --- | --- |
| **Section** | **Description** |
| Project Overview | The objective of this project is to build a machine learning model capable of identifying and classifying toxic comments on social media platforms. The focus is on reducing online harassment, promoting a healthy digital environment, and enabling moderation through automation. |
| Data Collection Plan | Data will be sourced from publicly available datasets, social media platforms' comment feeds via APIs, and crowdsourced annotations. Efforts will ensure the inclusion of diverse and balanced data across toxicity types, languages, and user demographics. |
| Raw Data Sources Identified | For the project on toxic comment classification for social media, various raw data sources have been identified to provide the foundation for training and evaluating the classification model. The |

|  |  |
| --- | --- |
|  | primary data source is **user-generated content**, including comments, replies, and posts extracted from social media platforms like Twitter, Reddit, and Facebook.  To enhance the dataset, **publicly available labeled datasets**, such as those from Kaggle (e.g., the Jigsaw Toxic Comment Classification dataset), are utilized, containing annotations for various categories like hate speech, obscenity, and threats. |

**Raw Data Sources Template**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source Name** | **Description** | **Location/URL** | **Format** | **Size** | **Access Permissions** |
| Jigsaw Toxic Comments | Labeled dataset for toxic comments classification. | Kaggle Dataset | CSV | 500  MB | Public |
| Reddit Comments | Comments from Reddit API annotated for toxicity. | Reddit API | JSON | 1 GB | Private (with access) |
| Twitter Comments | Real-time tweets collected and labeled manually. | Twitter API | CSV | 800  MB | API Key (Restricted Access) |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Crowdsource Data | User-annotated toxicity classifications. | Internal collection from surveys. | CSV | 100  MB | Private |