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 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 toxic comment classification project aims to automatically identify toxic language in comments, categorizing them as toxic, severe_toxic, obscene, threat, insult, and identity_hate. Using train.csv for labeled data, text comments were tokenized for model input. Four deep learning models (ANN, CNN, LSTM, BiLSTM) were trained. The final model was deployed in a Flask app with a user-friendly UI, allowing predictions on new comments. The results are saved in submission.csv for each toxicity category.
Data Collection Plan	1. Social Media: Collect comments from Twitter, Reddit, and YouTube, focusing on controversial topics where toxicity is common.

	2. News and Blog Comment Sections: Gather comments from news sites (e.g., CNN) and platforms like Medium, which often feature polarized discussions. 3. Online Forums: Use forums like Quora and gaming forums where toxic language can be prevalent in heated debates. 4. Existing Datasets: Supplement with labeled toxic comment datasets like the Jigsaw dataset and others on Kaggle.
Raw Data Sources Identified	1. Jigsaw/Conversation AI Toxicity Dataset ● Description: Contains comments from Wikipedia discussions, labeled for various toxicity types (e.g., toxic, obscene, threat). ● Data Type: Labeled text comments from Wikipedia. 2. Toxic Comment Classification Dataset - Description: A dataset with labeled toxic comments from various online platforms, categorized by types of toxicity. - Data Type: Labeled comments with binary toxicity labels.

Raw Data Sources Template

Source Name	Description	Location/URL	Format	Size	Access Permissions
Kaggle	A collection of comments from Wikipedia talk pages, labeled with toxicity categories	https://www.kaggle.com/c/jigsaw-toxic-comment-	CSV	~350 MB (~358, 400 KB or	Public

	such as toxic, severe_toxic, obscene, threat, insult, and identity_hate.	classification-challenge/data		0.35 GB)	