Data Intake Report

Name: Hate Speech Detection
Report date: 16 May 2025
Internship Batch: LISUM44
Version: 1.0
Data intake by: Yashraj Parmar
Data intake reviewer: Data Glacier
Data storage location: https://github.com/yparmar2024/Data-Glacier/tree/main/Week%207

Tabular data details:

test_tweets_anuFYb8.csv:

Total number of observations	16,130
Total number of files	1
Total number of features	2
Base format of the file	.csv
Size of the data	1.6 MB

train_E6oV3IV.csv:

Total number of observations	29,530
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	3.1 MB

Proposed Approach:

Taking a brief look at the data, we can see that for the test file, we have 2 columns: id and tweet. The id column has the id of the tweet, a unique identifier for each tweet to help identify the same tweet in both files. The tweet column has the actual text of the tweet, this is what is more relevant to us in the machine learning model. The train file has 3 columns: label, id, and tweet. The label contains the true value, either 0 or 1, where 0 represents no hate speech presented in text, and 1 represents hate speech presented in text. The id column represents the unique identifier for each tweet to help identify the same tweet in both files. The tweet column has the actual text of the tweet. We know what the data looks like, we can assume that the labels are true for each tweet since it's in the test file. There are no more assumptions to be made and the rest of the data preprocessing will be done in the next week.