Data Intake Report

Name: Hate Speech detection using Transformers

Report date: 09/01/23

Internship Batch: LISUM25

Version: 1.0

Data intake by: Faris Chaudhry

Data intake reviewer: N/A

Data storage location: <https://github.com/farischaudhry/twitter-sentiment-analysis>

**Tabular data details:**

test\_tweets.csv

|  |  |
| --- | --- |
| **Total number of observations** | 16130 |
| **Total number of files** | 1 |
| **Total number of features** | 2 |
| **Base format of the file** | .csv |
| **Size of the data** | 1.56 MB |

train.tweets.csv

|  |  |
| --- | --- |
| **Total number of observations** | 29530 |
| **Total number of files** | 1 |
| **Total number of features** | 3 |
| **Base format of the file** | .csv |
| **Size of the data** | 2.96 MB |

**Proposed Approach:**

Data in the training file is labelled whereas in the test file it is not. The ID gives a primary key for the data. Dedup validation can be performed on both files by removing any duplicated tweets or, where two different tweets have the same ID, changing the ID of one.

Duplication, null, and incorrect keys can easily be checked beforehand on Kaggle:

A screenshot of a computer

Description automatically generated

Assumptions:

* The data is derived from real tweets.
* The training data is labelled correctly.
* The training and test data are from the same domain.
* The amount of hate speech compared to non-hate speech reflects the proportion on the platform (see below). Approximately 6% of the labels are classified as hate speech.
* Tweets are below the limit of 200 characters and are formatted in the same way.

A screenshot of a computer

Description automatically generated