Data Intake Report

Name: Hate Speech Detection using Transformers (Deep Learning)
Report date: 19.08.2022
Internship Batch: LISUM11
Version: 1.0
Data intake by: Kristina Kaliagina
Data intake reviewer: Data Glacier
Data storage location:
https://www.kaggle.com/datasets/vkrahul/twitter-hate-speech?select=train_E6oV3lV.csv

Tabular data details:

Total number of observations	49159
Total number of files	2
Total number of features	5
Base format of the file	.csv
Size of the data	4,7MB

Proposed Approach:

- Identification:
    - 2 files:
        - train.csv:
            - id (int64)
            - label (int64)
            - tweet (object)
        - test.csv:
            - id (int64)
            - tweet (object)
    - New features will be added, such as "hashtag".

- Assumptions:
    - The variable "label" is a binary variable that takes 1 when the tweet refers to hate speech, 0 when it doesn't.
    - The variable "label" needs to be predicted for tweets from file “test.csv”.
    - The variable "tweet" contains the original tweets with noise.