Data Intake Report

Name: Hate Speech Detection using Transformers
Report date: Nov. 2023
Internship Batch: LISUM25
Version:<1.0>
Data intake by: Seoyoung Kim
Data intake reviewer:
Data storage location: https://www.kaggle.com/datasets/vkrahul/twitter-hate-speech?select=train_E6oV3lV.csv

Tabular data details:

1. train_E6oV3lV

Total number of observations	31962
Total number of files	1
Total number of features	3
Base format of the file	csv
Size of the data	3.1 MB

2. test_tweets_anuFYb8

Total number of observations	17197
Total number of files	1
Total number of features	2
Base format of the file	csv
Size of the data	1.6 MB

Proposed Approach:

- Mention approach of dedup validation (identification)

  1. Explore dataset to identify potential duplicates. Errors during data collection, preprocessing, or merging datasets can result in duplication.

  2. When possible, unique identifiers are used such as Tweet ID or User ID, to flag or remove exact duplicates.

  3. Implement a text-based deduplication approach to identify and process tweets containing identical or very similar text content. You can use techniques like the cosine similarity.

- Mention your assumptions (if you assume any other thing for data quality analysis)

1. Assume that the provided label (0 or 1) for hate speech is accurate and assigned based on a reliable annotation process.

2. Assume that the text_format properties of the training and test datasets follow a consistent format and that all variations are within acceptable ranges.

3. Assume that common types of noise in Twitter data (e.g. hashtags, mentions, emojis) have been properly handled during preprocessing.

4. Assume that the dataset is representative of real-world Twitter data in terms of language usage, topics, and demographics to ensure the generalizability of the model.

5. Assume that the training and testing datasets are consistent in terms of data distribution, noise, and characteristics, allowing the model to generalize well.