The Twitter data found in this Google Drive folder was collected under the pretense that the data could be used by INTRNLP students of DLSU (AY2021-2022 T1) for any academic requirement(s) related to the course. If you don’t fall under this category, please contact [edward.tighe@dlsu.edu.ph](mailto:edward.tighe@dlsu.edu.ph) to discuss a more ethical approach to utilizing the data. Tweet IDs are released publicly and can be recrawled if needed.

The data was collected by passing a list of tweet IDs (previously collected) through Twitter’s v2 API endpoint. Only the Tweet ID, author ID, created at (in UTC+8), language, and text were collected. The data is saved in CSV format and each CSV file has roughly 250,000 tweets (roughly 32mb per file).

The original list of tweet IDs came from a previous project that was originally focused on collecting data for studying COVID-related responses online. The collection method of the initial study focused on collecting tweets within the Philippines (via a bounding box based on the Philippine’s geographical location) and the collection period lasted from March 2020 to November 2020 – resulting in roughly 17.6 million tweets.

As the pretense of the original collection did not match with the goals of the current INTRNLP class, the original data could not be used directly. Hence, to create the current dataset, we had to recrawl the data given the tweet IDs to verify if the tweet was still “alive” (i.e. not deleted by the user and still in a public setting). This manner respects any users who deleted their data and/or changed the privacy settings of their account since the previous collection.

Currently (as of 9:40am, 1 Jan 2022), the tweets are still being crawled and the Google Drive folder only has 6 files. We are expecting only half of the original dataset size to be collectable, so hopefully we’ll end up with roughly 36 CSV files or about 9 million tweets. As more data is crawled, the folder and this readme will be updated accordingly.