# Format of the data:

The dataset consists of training examples encoded in a csv file. For each example, we have the `post` (the text that was tweeted), the major `class` (`not\_hate`, `implicit\_hate`, `explicit\_hate`), and the `implicit\_class` (`white\_grievance`, `incitement`, `inferiority`, `irony`, `stereotypes`, `other`). Only `implicit\_hate` examples have a non-`other` value for their `implicit\_class`

# Data examples:

| Post | Class | Implicit Class |

| --- | --- | --- |

|’here's how the white house would cut legal immigration in half via’ | not\_hate | N/A

|’you think if a white man tweeted the racist tweets did but vice versa would he still have a job ? #doublestandard’ | implicit\_hate | white\_grievance

# Source of data:

We started with the Implicit Hate Speech dataset from ["Latent Hatred: A Benchmark for Understanding Implicit Hate Speech"](<https://github.com/SALT-NLP/implicit-hate>), and made minimal preprocessing modifications. In particular, we ignore any `stg3` files, because the determining target and implied statement of implicit hate is beyond the scope of our project. Additionally, we join the `stg1` and `stg2` files to create a dataset of posts, where each post has both a class and an implicit class (which is NaN for `not\_hate` and `explicit\_hate`). We also removed a few duplicate `post`s.

# Split details:

Train:16436 rows

Dev:2055 rows

Test: 2989 rows

Some implicit hate posts did not have an implicit class, and the implicit class `other` ended up being very small. We put these examples in the test set for ease later on when working with the data, which is why the test set is a bit larger than the dev set