# Week 8 Deliverables

# Group Details


- Name: William Harvey
- Email: will.harvey@att.net
- Country: United States of America
- College: University of California, Berkeley
- Specialization: NLP Analyst

Week 8: Deliverables | Untitled.ipynb (auto-G) - Jupyt

data100.datahub.berkeley.edu/user/wsharvey/lab/workspaces/auto-G/tree/DataGlacier/Final%20Proj/Untitled.ipynb

File Edit View Run Kernel Tabs Settings Help

Untitled.ipynb | train.csv | test_tweets_anuFYb8.csv

Code | Interface | Python 3 (ipykernel)

[1]:
```
import pandas as pd
import numpy as np

test_tweets = pd.read_csv("Raw Data/test_tweets_anuFYb8.csv")
train_tweets = pd.read_csv("Raw Data/train.csv")
```

The types of data per column per dataset can be seen below.

[7]:
```
for col in train_tweets.columns:
    print(f'Column "{col}" in training data has type {type(col)}')

for col in test_tweets.columns:
    print(f'Column "{col}" in test data has type {type(col)}')
```

```
Column "id" in training data has type <class 'str'>
Column "label" in training data has type <class 'str'>
Column "tweet" in training data has type <class 'str'>
Column "id" in test data has type <class 'str'>
Column "tweet" in test data has type <class 'str'>
```

[4]:
```
nan_vals = len(train_tweets[train_tweets['tweet'].isna()])
labeled_1 = len(train_tweets[train_tweets['label'] == 1])
labeled_0 = len(train_tweets[train_tweets['label'] == 0])

print(f"There are {nan_vals} NaN values in the data")
print(f"There are {labeled_0} rows that are NOT classified as hate speech, while there are {labeled_1} rows that are classified as hate speech, leading to imbal
```

```
There are 0 NaN values in the data
There are 29720 rows that are NOT classified as hate speech, while there are 2242 rows that are classified as hate speech, leading to imbalanced data
```

In order to account for this data imbalance, I will use the "resample" module of sklearn.utils, in order to create a balanced data set. This must be done because there are so little samples where hate speech is classified, so the model will ultimately favor classifying new tweets as NOT hate speech because there is such an overwhelming amount of data for this aspect. Since there are no NaN values or outliers, no further data processing has to be done to account for this area.

[ ]:

Simple | 0 | 2 | Python 3 (ipykernel) | Idle | Mem: 295.56 / 2048.00 MB | Mode: Command | Ln 6, Col 173 | Untitled.ipynb