Group Name: AO
Name: Alaa Eddine Osta
Email: ostaalaaeddine@gmail.com
Country: Lebanon
College: Üsküdar University
Specialization: NLP
Project Deadline: 11/25/2021

Problem Description

The term hate speech is any form of expression (verbal, written, or behavioral communication) that attacks or uses derogatory or discriminatory language against a person or group. This hate could be towards religion, ethnicity, nationality, race, color, ancestry, sex, or other identity factors. In this problem, we need to build a machine learning model that predicts which tweets have hate speech in them.

Hate Speech Detection is generally a task of sentiment classification. To classify hate speech from a piece of text, we need to train the model on data used to classify sentiments. So for the task of the hate speech detection model, we will use Twitter tweets to identify tweets containing hate speech.

Data Understanding

	id	label	tweet
0	1	0	@user when a father is dysfunctional and is s...
1	2	0	@user @user thanks for #lyft credit i can't us...
2	3	0	bihday your majesty
3	4	0	#model i love u take with u all the time in ...
4	5	0	factsguide: society now #motivation
...	...	...	...
31957	31958	0	ate @user isz that youuu?ð□□□ð□□□ð□□□ð□□□ð□□□ð...
31958	31959	0	to see nina turner on the airwaves trying to...
31959	31960	0	listening to sad songs on a monday morning otw...
31960	31961	1	@user #sikh #temple vandalised in in #calgary,...
31961	31962	0	thank you @user for you follow

31962 rows × 3 columns

Figure 1: Data Frame of the Data

Observations from Figure 1:

- There is a lot of noise in the data ()
- There are three total columns in the dataset
- The number of observations is equal to 31962

```
RangeIndex: 31962 entries, 0 to 31961
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   id      31962 non-null  int64
 1   label   31962 non-null  int64
 2   tweet   31962 non-null  object
dtypes: int64(2), object(1)
memory usage: 749.2+ KB
```

Figure 2: Data Frame Information

Observations from Figure 2:

- There are no null values in the dataset
- The data type of the tweet column is String object
- The data type of both id and label columns is int64
- The size of the data is 749.2+ KB

	label	count	percentage
0	0	29720	92.98542
1	1	2242	7.01458

Figure 3: The number of examples each class has

Observations from Figure 3:

- The data given is unbalanced where the number of data samples labeled 0 is about 13 times higher than the samples labeled 1, which causes an overfitting problem for models, as they mostly predict zeros.

To overcome the data set problems we observed we can consider the following approaches:
- Cleaning data (removing noise from data)
- Transforming data (transforming data into numerical values)
- Using pre-trained vectorizers for data transformation
- Using a pre-trained Transformer and fine tuning it to the given data