DataGlacier NLP Interns Group

MEMBER DETAILS

- Member 1:
  - Name – Gabriel Aluede
  - Email – ghabyaluede@gmail.com
  - Country – United Kingdom
  - College – Teesside University
  - Specialization – NLP
- Member 2:
  - Name – Yosha Udaya Shriyan
  - Email – yosha.shriyan@gmail.com
  - Country – United Kingdom
  - College – Royal Holloway University of London
  - Specialization – NLP

PROBLEM DESCRIPTION

The term hate speech is understood as any type of verbal, written or behavioural communication that attacks or uses derogatory or discriminatory language against a person or group based on what they are, in other words, based on their religion, ethnicity, nationality, race, colour, ancestry, sex or another identity factor. In this problem, we will take you through a hate speech detection model with Machine Learning and Python.

DATA UNDERSTANDING

Data Attributes

Label : 0 or 1
Text_Format : original tweets with noise
Source: https://www.kaggle.com/datasets/vkrahul/twitter-hate speech?select=train_E6oV3lV.csv
NA Values: None

```
train_data.isnull().sum()
```

```
id       0
label    0
tweet    0
dtype: int64
```

Train_Data

```
[ ]  #A quick peep at the data
     train_data.head()
```

	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

Test_Data

```
[ ]  test_data.head()
```

	id	tweet
0	31963	#studiolife #aislife #requires #passion #dedic...
1	31964	@user #white #supremacists want everyone to s...
2	31965	safe ways to heal your #acne!! #altwaystohe...
3	31966	is the hp and the cursed child book up for res...
4	31967	3rd #bihday to my amazing, hilarious #nephew...

Data Size

```
sum(train_data["label"]==1)
```

```
2242
```

```
sum(train_data["label"]==0)
```

```
29720
```

```
[ ]  train_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
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
```

```
[ ]  train_data.describe()
```

	id	label
count	31962.000000	31962.000000
mean	15981.500000	0.070146
std	9226.778988	0.255397
min	1.000000	0.000000
25%	7991.250000	0.000000
50%	15981.500000	0.000000
75%	23971.750000	0.000000
max	31962.000000	1.000000

GITHUB REPO LINK

https://github.com/YoshaRHUL/DataGlacierProject