```
# This Python 3 environment comes with many helpful analytics libraries installed
# It is defined by the kaggle/python Docker image: https://github.com/kaggle/docker-python
# For example, here's several helpful packages to load

import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

# Input data files are available in the read-only "../input/" directory
# For example, running this (by clicking run or pressing Shift+Enter) will list all files under the input directory

import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

# You can write up to 20GB to the current directory (/kaggle/working/) that gets preserved as output when you create a version using "Sa
# You can also write temporary files to /kaggle/temp/, but they won't be saved outside of the current session


import pandas as pd
pd.set_option('display.width', None)  # Auto-adjust width
pd.set_option('display.max_colwidth', None)  # No limit on column width


from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```
df_train = pd.read_csv("/content/drive/MyDrive/datasets_project/train_E6oV3lV.csv"
, header=0)
df_train.head(20)
```

	id	label	tweet
0	1	0	@user when a father is dysfunctional and is so selfish he drags his kids into his dysfunction. #run
1	2	0	@user @user thanks for #lyft credit i can't use cause they don't offer wheelchair vans in pdx. #disapointed #getthanked
2	3	0	bihday your majesty
3	4	0	#model i love u take with u all the time in urð▯▯▯±!!! ð▯▯▯ð▯▯▯ð▯▯▯ð▯▯▯ð▯▯▯¦ð▯▯▯¦ð▯▯▯¦
4	5	0	factsguide: society now #motivation
5	6	0	[2/2] huge fan fare and big talking before they leave. chaos and pay disputes when they get there. #allshowandnogo
6	7	0	@user camping tomorrow @user @user @user @user @user @user @user dannyâ▯▯¦
7	8	0	the next school year is the year for exams.ð▯▯¯ can't think about that ð▯▯ #school #exams #hate #imagine #actorslife #revolutionschool #girl
8	9	0	we won!!! love the land!!! #allin #cavs #champions #cleveland #clevelandcavaliers â▯▯¦
9	10	0	@user @user welcome here ! i'm it's so #gr8 !
10	11	0	â▯▯ #ireland consumer price index (mom) climbed from previous 0.2% to 0.5% in may #blog #silver #gold #forex
11	12	0	we are so selfish. #orlando #standwithorlando #pulseshooting #orlandoshooting #biggerproblems #selfish #heabreaking #values #love #
12	13	0	i get to see my daddy today!! #80days #gettingfed
13	14	1	@user #cnn calls #michigan middle school 'build the wall' chant " #tcot
14	15	1	no comment! in #australia #opkillingbay #seashepherd #helpcovedolphins #thecove #helpcovedolphins
15	16	0	ouch...junior is angryð▯▯▯#got7 #junior #yugyoem #omg
16	17	0	i am thankful for having a paner. #thankful #positive
17	18	1	retweet if you agree!
18	19	0	its #friday! ð▯▯▯ smiles all around via ig user: @user #cookies make people

Next steps: View recommended plots New interactive sheet

```
df_train[df_train["label"] == 0]["tweet"].iloc[2]
```

```
' bihday your majesty'
```

```
df_train.shape
```

```
(31962, 3)
```

```
df_train[df_train["label"] == 0].count()
```

	0
id	29720
label	29720
tweet	29720

dtype: int64

```
#df.query('label == 0').count()
len(df_train[df_train["label"] == 0])
```

```
29720
```

```
len(df_train[df_train["label"] == 1])
```

```
2242
```

```
df_test = pd.read_csv("/content/drive/MyDrive/datasets_project/test_tweets_anuFYb8.csv")
```

```
df_test.shape
```

```
(17197, 2)
```

```
df_test.head()
```

	id	tweet
0	31963	#studiolife #aislife #requires #passion #dedication #willpower to find #newmaterialsâ¦
1	31964	@user #white #supremacists want everyone to see the new â¦ #birdsâ¦ #movie â¦ and hereâ¦s why
2	31965	safe ways to heal your #acne!! #altwaystoheal #healthy #healing!!
3	31966	is the hp and the cursed child book up for reservations already? if yes, where? if no, when? ð¤¤ð¤¤ð¤¤ #harrypotter #pottermore #favorite
4	31967	3rd #bihday to my amazing, hilarious #nephew eli ahmir! uncle dave loves you and missesâ¦

Next steps: View recommended plots New interactive sheet

```
from matplotlib import pyplot as plt
import seaborn as sns
```

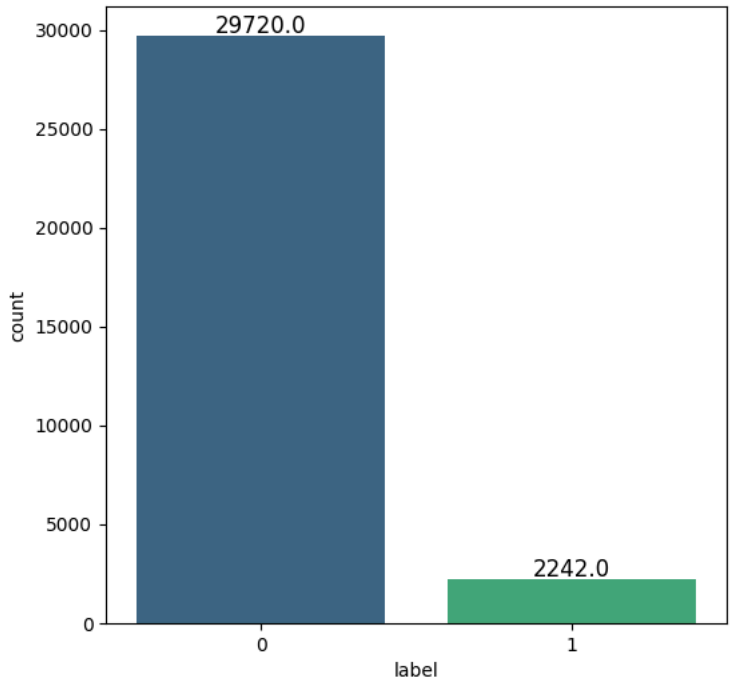
```
plt.figure(figsize=(6,6))
ax = sns.countplot(data=df_train, x="label", palette='viridis')
for p in ax.patches:
    ax.annotate(f'{p.get_height()}',
                (p.get_x() + p.get_width() / 2., p.get_height()),
                ha='center', va='center',
                fontsize=12, color='black',
                xytext=(0, 5),  # Adjust text position (optional)
                textcoords='offset points')

# Show the plot
plt.show()
```

```
<ipython-input-15-f27fc25f6679>:2: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `le

  ax = sns.countplot(data=df_train, x="label", palette='viridis')
```




```
df_train.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

```
df_train.groupby('label').describe()
```

	id							
	count	mean	std	min	25%	50%	75%	max
label								
0	29720.0	15974.454441	9223.783469	1.0	7981.75	15971.5	23965.25	31962.0
1	2242.0	16074.896075	9267.955758	14.0	8075.25	16095.0	24022.00	31961.0

```
df_train.head(20)
```

	id	label	tweet
0	1	0	@user when a father is dysfunctional and is so selfish he drags his kids into his dysfunction. #run
1	2	0	@user @user thanks for #lyft credit i can't use cause they don't offer wheelchair vans in pdx. #disapointed #getthanked
2	3	0	bihday your majesty
3	4	0	#model i love u take with u all the time in urð▯▯±!!! ð▯▯▯ð▯▯▯ð▯▯▯ð▯▯▯ð▯▯¦ð▯▯¦ð▯▯¦
4	5	0	factsguide: society now #motivation
5	6	0	[2/2] huge fan fare and big talking before they leave. chaos and pay disputes when they get there. #allshowandnogo
6	7	0	@user camping tomorrow @user @user @user @user @user @user @user dannyâ▯¦
7	8	0	the next school year is the year for exams.ð▯▯¯ can't think about that ð▯▯ #school #exams #hate #imagine #actorslife #revolutionschool #girl
8	9	0	we won!!! love the land!!! #allin #cavs #champions #cleveland #clevelandcavaliers â▯¦
9	10	0	@user @user welcome here ! i'm it's so #gr8 !
10	11	0	â▯▯ #ireland consumer price index (mom) climbed from previous 0.2% to 0.5% in may #blog #silver #gold #forex
11	12	0	we are so selfish. #orlando #standwithorlando #pulseshooting #orlandoshooting #biggerproblems #selfish #heabreaking #values #love #
12	13	0	i get to see my daddy today!! #80days #gettingfed
13	14	1	@user #cnn calls #michigan middle school 'build the wall' chant " #tcot
14	15	1	no comment! in #australia #opkillingbay #seashepherd #helpcovedolphins #thecove #helpcovedolphins
15	16	0	ouch...junior is angryð▯▯▯#got7 #junior #yugyoem #omg
16	17	0	i am thankful for having a paner. #thankful #positive
17	18	1	retweet if you agree!
18	19	0	its #friday! ð▯▯▯ smiles all around via ig user: @user #cookies make people

Next steps: View recommended plots New interactive sheet

Cleanse Data

```
import re
```

```
def clean_text(text):
    text = re.sub(r"http\S+", "", text) # remove URLs
    text = re.sub(r"@\w+", "", text) #remove mentions
    text = re.sub(r"#\w+", "", text) #remove hashtags
    #text = re.sub(r"[^\w\s]", "", text) #remove special characters
    return text.lower().strip()
```

```
df_train['tweet'][3]
```

```
'#model   i love u take with u all the time in urð\x9f\x93+!!! ð\x9f\x98\x99ð\x9f\x98\x8eð\x9f\x91\x84ð\x9f\x91\x85ð\x9f\x92'ð\x9f
```

```
df_train['cleaned_tweet']= df_train['tweet'].apply(clean_text)
```

```
df_train.head(20)
```

	id	label	tweet	cleaned_tweet
0	1	0	@user when a father is dysfunctional and is so selfish he drags his kids into his dysfunction. #run	when a father is dysfunctional and is so selfish he drags his kids into his dysfunction.
1	2	0	@user @user thanks for #lyft credit i can't use cause they don't offer wheelchair vans in pdx. #disapointed #getthanked	thanks for credit i can't use cause they don't offer wheelchair vans in pdx.
2	3	0	bihday your majesty	bihday your majesty
3	4	0	#model i love u take with u all the time in urð▯▯±!!! ð▯▯▯ð▯▯▯ð▯▯▯ð▯▯▯ð▯▯¦ð▯▯¦ð▯▯¦	i love u take with u all the time in urð▯▯±!!! ð▯▯▯ð▯▯▯ð▯▯▯ð▯▯▯ð▯▯¦ð▯▯¦ð▯▯¦
4	5	0	factsguide: society now #motivation	factsguide: society now
5	6	0	[2/2] huge fan fare and big talking before they leave. chaos and pay disputes when they get there. #allshowandnogo	[2/2] huge fan fare and big talking before they leave. chaos and pay disputes when they get there.
6	7	0	@user camping tomorrow @user @user @user @user @user @user @user dannyâ▯¦	camping tomorrow dannyâ▯¦
7	8	0	the next school year is the year for exams.ð▯▯¯ can't think about that ð▯▯- #school #exams #hate #imagine #actorslife #revolutionschool #girl	the next school year is the year for exams.ð▯▯¯ can't think about that ð▯▯
8	9	0	we won!!! love the land!!! #allin #cavs #champions #cleveland #clevelandcavaliers â▯¦	we won!!! love the land!!! â▯¦
9	10	0	@user @user welcome here ! i'm it's so #gr8 !	welcome here ! i'm it's so !
10	11	0	â▯▯ #ireland consumer price index (mom) climbed from previous 0.2% to 0.5% in may #blog #silver #gold #forex	â▯▯ consumer price index (mom) climbed from previous 0.2% to 0.5% in may
11	12	0	we are so selfish. #orlando #standwithorlando #pulseshooting #orlandoshooting #biggerproblems #selfish #heabreaking #values #love #	we are so selfish. #
12	13	0	i get to see my daddy today!! #80days #gettingfed	i get to see my daddy today!!
13	14	1	@user #cnn calls #michigan middle school 'build the wall' chant " #tcot	calls middle school 'build the wall' chant "
14	15	1	no comment! in #australia #opkillingbay #seashepherd #helpcovedolphins #thecove #helpcovedolphins	no comment! in
15	16	0	ouch...junior is angryð▯▯▯#got7 #junior #yugyoem #omg	ouch...junior is angryð▯▯▯
16	17	0	i am thankful for having a paner. #thankful #positive	i am thankful for having a paner.

Next steps: View recommended plots New interactive sheet

```
def decode_text(text):
    try:
        #Handling the characters's error
        return text.encode('latin1').decode('utf-8', errors='ignore')
    except UnicodeDecodeError:
        return text


def contains_invalid_encoding(text):
    # Kiểm tra nếu text là chuỗi và tìm ký tự không hợp lệ
    if isinstance(text, str):
        # Regex tìm các ký tự không phải Unicode hợp lệ (biểu hiện của mã hóa sai)
        return bool(re.search(r'[^\x00-\x7F]+', text))  # Tìm ký tự ngoài bảng mã ASCII chuẩn
    return False  # Nếu không phải chuỗi, trả về False


# Đảm bảo cột 'tweet' là kiểu chuỗi trước khi áp dụng hàm kiểm tra
df_train['cleaned_tweet'] = df_train['cleaned_tweet'].astype(str)

# Tạo một cột mới để lưu tweet đã được mã hóa lại
df_train['decoded_tweet'] = df_train['cleaned_tweet'].apply(lambda x: decode_text(x) if contains_invalid_encoding(x) else x)

# Lọc ra những dòng chứa ký tự mã hóa sai (ngoài bảng ASCII)
invalid_tweets = df_train[df_train['cleaned_tweet'].apply(contains_invalid_encoding)]

# Kiểm tra các tweet đã được mã hóa lại
invalid_tweets_after_decoding = df_train[df_train['cleaned_tweet'] != df_train['decoded_tweet']]

# Hiển thị những dòng chứa ký tự mã hóa sai
print(f"Những tweet chứa ký tự mã hóa sai (ngoài bảng ASCII):")
print(invalid_tweets[['cleaned_tweet']])  # Hiển thị chỉ cột tweet
```

```
Những tweet chứa ký tự mã hóa sai (ngoài bảng ASCII):
                                                cleaned_tweet
3      i love u take with u all the time in urð▯▯±!!! ð▯▯▯ð▯▯▯ð▯▯▯ð▯▯¦ð▯▯¦ð▯▯¦
6                                   camping tomorrow        dannyâ▯¦
7      the next school year is the year for exams.ð▯▯¯ can't think about that ð▯▯
8                                we won!!! love the land!!!       â▯¦
10       â▯▯  consumer price index (mom) climbed from previous 0.2% to 0.5% in may
...                                                                           ...
```

```
31939                                   whew  ð□□□ it's a productive and   !!!
31942                                         this week is flying by    -   □¦
31945                                                feel like... ð□□□ð□□¶ð□□□
31955                                 less than 2 weeks ð□□□ð□□□ð□□¼ð□□¹ð□□□ð□□µ
31957                ate  isz that youuu?ð□□□ð□□□ð□□□ð□□□ð□□□ð□□□ð□□□ð□□□â□¤ï¸□

[10751 rows x 1 columns]
```

```
print(f"Những tweet đã được mã hóa lại:")
print(invalid_tweets_after_decoding[['cleaned_tweet', 'decoded_tweet']])
```

```
Những tweet đã được mã hóa lại:
                                                          cleaned_tweet  \
3      i love u take with u all the time in urð□□±!!! ð□□□ð□□□ð□□□ð□□□¦ð□□¦ð□□¦
6                                            camping tomorrow        dannyâ□¦
7      the next school year is the year for exams.ð□□¯ can't think about that ð□□
8                                              we won!!! love the land!!!       â□¦
10       â□□  consumer price index (mom) climbed from previous 0.2% to 0.5% in may
...                                                                            ...
31939                                   whew  ð□□□ it's a productive and   !!!
31942                                         this week is flying by    -   □¦
31945                                                feel like... ð□□□ð□□¶ð□□□
31955                                 less than 2 weeks ð□□□ð□□□ð□□¼ð□□¹ð□□□ð□□µ
31957                ate  isz that youuu?ð□□□ð□□□ð□□□ð□□□ð□□□ð□□□ð□□□ð□□□â□¤ï¸□

                                                        decoded_tweet
3                        i love u take with u all the time in ur🚉!!! 😙😎👄👅💦💦💦
6                                            camping tomorrow        danny…
7        the next school year is the year for exams.😳 can't think about that 😭
8                                              we won!!! love the land!!!       …
10     ~  consumer price index (mom) climbed from previous 0.2% to 0.5% in may
...                                                                          ...
31939                                   whew  😅 it's a productive and   !!!
31942                                         this week is flying by    -
31945                                                feel like... 😝🐶😎
31955                                                 less than 2 weeks 😅🙏🍹😎🎵
31957                                             ate  isz that youuu?😍😍😍😍😍😍😍😍😍❤️

[10751 rows x 2 columns]
```

```
repeated_char_train = []
for i, tweet in enumerate(df_train['decoded_tweet']):
  if re.match(r'\w*(\w)\1+',tweet):
    repeated_char_train.append(i)
```

```
len(repeated_char_train)
```

```
4583
```

```
#Repeated char function
def repeated_char(text):
  text = re.sub(r'(\w)\1{2,}',r'\1',text)
  return text
```

```
df_train['clean_duplicate']=df_train['decoded_tweet'].apply(repeated_char)
```

```
df_train.head(20)
```

	id	label	tweet	cleaned_tweet	decoded_tweet	clean_duplicate
0	1	0	@user when a father is dysfunctional and is so selfish he drags his kids into his dysfunction. #run	when a father is dysfunctional and is so selfish he drags his kids into his dysfunction.	when a father is dysfunctional and is so selfish he drags his kids into his dysfunction.	when a father is dysfunctional and is so selfish he drags his kids into his dysfunction.
1	2	0	@user @user thanks for #lyft credit i can't use cause they don't offer wheelchair vans in pdx. #disapointed #getthanked	thanks for credit i can't use cause they don't offer wheelchair vans in pdx.	thanks for credit i can't use cause they don't offer wheelchair vans in pdx.	thanks for credit i can't use cause they don't offer wheelchair vans in pdx.
2	3	0	bihday your majesty	bihday your majesty	bihday your majesty	bihday your majesty
3	4	0	#model i love u take with u all the time in urð▯▯±!!! ð▯▯▯ð▯▯▯ð▯▯▯ð▯▯▯ð▯▯▯ð▯▯▯ð▯▯▯	i love u take with u all the time in urð▯▯±!!! ð▯▯▯ð▯▯▯ð▯▯▯ð▯▯▯ð▯▯▯ð▯▯▯	i love u take with u all the time in ur 📱 !!! 😙 😎 👄 👅 💦 💦 💦	i love u take with u all the time in ur 📱 !!! 😙 😎 👄 👅 💦 💦 💦
4	5	0	factsguide: society now #motivation	factsguide: society now	factsguide: society now	factsguide: society now
5	6	0	[2/2] huge fan fare and big talking before they leave. chaos and pay disputes when they get there. #allshowandnogo	[2/2] huge fan fare and big talking before they leave. chaos and pay disputes when they get there.	[2/2] huge fan fare and big talking before they leave. chaos and pay disputes when they get there.	[2/2] huge fan fare and big talking before they leave. chaos and pay disputes when they get there.
6	7	0	@user camping tomorrow @user @user @user @user @user @user @user dannyâ▯¦	camping tomorrow dannyâ▯¦	camping tomorrow danny…	camping tomorrow danny…
7	8	0	the next school year is the year for exams.ð▯▯¯ can't think about that ð▯▯- #school #exams #hate #imagine #actorslife #revolutionschool #girl	the next school year is the year for exams.ð▯▯¯ can't think about that ð▯▯	the next school year is the year for exams. 😱 can't think about that 😭	the next school year is the year for exams. 😱 can't think about that 😭
8	9	0	we won!!! love the land!!! #allin #cavs #champions #cleveland #clevelandcavaliers â▯¦	we won!!! love the land!!! â▯¦	we won!!! love the land!!! …	we won!!! love the land!!! …
9	10	0	@user @user welcome here ! i'm it's so #gr8 !	welcome here ! i'm it's so !	welcome here ! i'm it's so !	welcome here ! i'm it's so !
10	11	0	â▯▯ #ireland consumer price index (mom) climbed from previous 0.2% to 0.5% in may #blog #silver #gold #forex	â▯▯ consumer price index (mom) climbed from previous 0.2% to 0.5% in may	~ consumer price index (mom) climbed from previous 0.2% to 0.5% in may	~ consumer price index (mom) climbed from previous 0.2% to 0.5% in may
11	12	0	we are so selfish. #orlando #standwithorlando #pulseshooting #orlandoshooting #biggerproblems #selfish #heabreaking #values #love #	we are so selfish. #	we are so selfish. #	we are so selfish. #
12	13	0	i get to see my daddy today!! #80days #gettingfed	i get to see my daddy today!!	i get to see my daddy today!!	i get to see my daddy today!!
13	14	1	@user #cnn calls #michigan middle school 'build the wall' chant " #tcot	calls middle school 'build the wall' chant "	calls middle school 'build the wall' chant "	calls middle school 'build the wall' chant "
14	15	1	no comment! in #australia #opkillingbay #seashepherd	no comment! in	no comment! in	no comment! in

Next steps: View recommended plots New interactive sheet

Emoji Processing

```
!pip install emoji
```

```
Collecting emoji
  Downloading emoji-2.14.0-py3-none-any.whl.metadata (5.7 kB)
Downloading emoji-2.14.0-py3-none-any.whl (586 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 586.9/586.9 kB 15.6 MB/s eta 0:00:00
Installing collected packages: emoji
Successfully installed emoji-2.14.0
```

```
import emoji
def emoji_text_trans(text):
    text= emoji.demojize(text).replace (':', ' ')

    #delete repeated emoji
    tokenize= text.split()
    repeated_list=[]
    for word in tokenize:
        if word not in repeated_list:
            repeated_list.append(word)

    text=' '.join(word for word in repeated_list)
    text= text.replace("_", " ").replace("-", " ")
    return text
```

```
df_train["deemoji_tweet"] = df_train["clean_duplicate"].apply(emoji_text_trans)
```

```
from sklearn.model_selection import train_test_split
#label
y= df_train.label
#features
x=df_train.deemoji_tweet
#split into test and train dataset with test size 20%
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2)
x_train.head()
```

	deemoji_tweet
8321	wedding teaser d & c ! photography by
11048	when your gingerjew friend won't text chu.
9957	tv on no breakfast. what's happening? frowning face
9774	can't wait for the uk or french tour of mark, matt and travis are killing it
21160	one of the biggest aims from these retreats is true happiness with being you

dtype: object

Cleansing df_test

Distill BERT

```
import torch
from transformers import DistilBertTokenizer, DistilBertForSequenceClassification
from sklearn.utils.class_weight import compute_class_weight
from torch.utils.data import DataLoader, TensorDataset
from sklearn.metrics import classification_report, confusion_matrix
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
```

```
# Initial Setup
device = torch.device('cuda' if torch.cuda.is_available() else 'cpu')
tokenizer = DistilBertTokenizer.from_pretrained('distilbert-base-uncased')

# Tokenization and Tensor Creation
texts = x_train.tolist()
labels = y_train.tolist()
encoded_inputs = tokenizer(
    texts,
    padding=True,
    truncation=True,
    max_length=128,
    return_tensors="pt"
)
input_ids = encoded_inputs['input_ids']
attention_mask = encoded_inputs['attention_mask']
labels = torch.tensor(labels)
```

```
/usr/local/lib/python3.10/dist-packages/huggingface_hub/utils/_auth.py:94: UserWarning:
The secret `HF_TOKEN` does not exist in your Colab secrets.
To authenticate with the Hugging Face Hub, create a token in your settings tab (https://huggingface.co/settings/tokens), set it as s
You will be able to reuse this secret in all of your notebooks.
Please note that authentication is recommended but still optional to access public models or datasets.
  warnings.warn(
tokenizer_config.json: 100%  48.0/48.0 [00:00<00:00, 653B/s]
vocab.txt: 100%  232k/232k [00:00<00:00, 3.78MB/s]
tokenizer.json: 100%  466k/466k [00:00<00:00, 3.00MB/s]
config.json: 100%  483/483 [00:00<00:00, 25.4kB/s]
```

```
# Dataset y DataLoader
dataset = TensorDataset(input_ids, attention_mask, labels)
train_loader = DataLoader(dataset, batch_size=16, shuffle=True)
```

```
# Model
model = DistilBertForSequenceClassification.from_pretrained('distilbert-base-uncased', num_labels=2)
model.to(device)

# Class Weighting to Handle Imbalance
class_weights = compute_class_weight('balanced', classes=np.unique(labels.numpy()), y=labels.numpy())
class_weights = torch.tensor(class_weights, dtype=torch.float).to(device)
criterion = torch.nn.CrossEntropyLoss(weight=class_weights)
optimizer = torch.optim.AdamW(model.parameters(), lr=5e-5)
```

```
model.safetensors: 100%                                    268M/268M [00:01<00:00, 217MB/s]
Some weights of DistilBertForSequenceClassification were not initialized from the model checkpoint at distilbert-base-uncased and ar
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
```

```
# Training
epochs = 5
training_loss = []
model.train()
for epoch in range(epochs):
    epoch_loss = 0
    for batch in train_loader:
        b_input_ids, b_attention_mask, b_labels = tuple(t.to(device) for t in batch)

        optimizer.zero_grad()
        outputs = model(
            input_ids=b_input_ids,
            attention_mask=b_attention_mask,
            labels=b_labels
        )
        loss = outputs.loss
        loss.backward()
        optimizer.step()
        epoch_loss += loss.item()

    avg_loss = epoch_loss / len(train_loader)
    training_loss.append(avg_loss)
    print(f"Epoch {epoch + 1}/{epochs}, Loss: {avg_loss:.4f}")

    save_path = "/content/drive/MyDrive/datasets_project"
    # Save the Model After Each Epoch
    model.save_pretrained(os.path.join(save_path, f"distilbert_model_epoch_{epoch + 1}"))
    tokenizer.save_pretrained(os.path.join(save_path, f"distilbert_model_epoch_{epoch + 1}"))
```

```
Epoch 1/5, Loss: 0.1662
Epoch 2/5, Loss: 0.0881
Epoch 3/5, Loss: 0.0410
Epoch 4/5, Loss: 0.0231
Epoch 5/5, Loss: 0.0172
```

```
test_dataset = TensorDataset(input_ids, attention_mask)
test_loader = DataLoader(test_dataset, batch_size=8)

# Perform Batch Inference
y_pred = []
model.eval()
with torch.no_grad():
    for batch in test_loader:
        b_input_ids, b_attention_mask = tuple(t.to(device) for t in batch)
        outputs = model(input_ids=b_input_ids, attention_mask=b_attention_mask)
        logits = outputs.logits
        y_pred.extend(torch.argmax(logits, dim=1).cpu().numpy())


y_pred = y_pred[:len(y_test)]

# Classification Report
print("\nClassification Report:\n")
print(classification_report(y_test, y_pred, target_names=['No Hate', 'Hate']))
```

```
Classification Report:

              precision    recall  f1-score   support

     No Hate       0.93      0.93      0.93      5916
        Hate       0.08      0.07      0.07       477

    accuracy                           0.87      6393
   macro avg       0.50      0.50      0.50      6393
weighted avg       0.86      0.87      0.87      6393
```

```
# Confusion Matrix
conf_matrix = confusion_matrix(y_test, y_pred)
plt.figure(figsize=(8, 6))
sns.heatmap(conf_matrix, annot=True, fmt='d', cmap='Blues', xticklabels=['No Hate', 'Hate'], yticklabels=['No Hate', 'Hate'])
plt.title('Confusion Matrix')
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.show()
```