Dummy

April 8, 2023

[1]:
```
from google.colab import drive
drive.mount('/content/drive')
```

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

[2]:
```
%cd /content/drive/MyDrive/Notebooks/hate_speech/
```

```
/content/drive/MyDrive/Notebooks/hate_speech
```

1 Hate Speech Project

1.1 Table of Contents

Introduction

Hypothesis Generation

Import Packages and Loading Data

Featurization

Introduction

1.1.1 Project 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.

Hate Speech Detection is generally a task of sentiment classification. So for training, a model that can classify hate speech from a certain piece of text can be achieved by training it on a data that is generally used to classify sentiments. So for the task of hate speech detection model, We will use the Twitter tweets to identify tweets containing Hate speech.

[ ]:
```
!pip3 uninstall --yes torch  torchtext
!pip3 install torch torchtext
```

```
Found existing installation: torch 1.13.1+cu116
Uninstalling torch-1.13.1+cu116:
  Successfully uninstalled torch-1.13.1+cu116
```

```
Found existing installation: torchtext 0.14.1
Uninstalling torchtext-0.14.1:
  Successfully uninstalled torchtext-0.14.1
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-
wheels/public/simple/
Collecting torch
  Downloading torch-2.0.0-cp39-cp39-manylinux1_x86_64.whl (619.9 MB)
                             619.9/619.9
MB 2.3 MB/s eta 0:00:00
Collecting torchdata
  Downloading
torchdata-0.6.0-cp39-cp39-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (4.6
MB)
                             4.6/4.6 MB
85.5 MB/s eta 0:00:00
Collecting nvidia-cuda-nvrtc-cu11==11.7.99
  Downloading nvidia_cuda_nvrtc_cu11-11.7.99-2-py3-none-manylinux1_x86_64.whl
(21.0 MB)
                             21.0/21.0 MB
66.8 MB/s eta 0:00:00
Collecting nvidia-cusparse-cu11==11.7.4.91
  Downloading nvidia_cusparse_cu11-11.7.4.91-py3-none-manylinux1_x86_64.whl
(173.2 MB)
                             173.2/173.2
MB 7.2 MB/s eta 0:00:00
Collecting nvidia-cudnn-cu11==8.5.0.96
  Downloading nvidia_cudnn_cu11-8.5.0.96-2-py3-none-manylinux1_x86_64.whl (557.1
MB)
                             557.1/557.1
MB 2.6 MB/s eta 0:00:00
Requirement already satisfied: typing-extensions in
/usr/local/lib/python3.9/dist-packages (from torch) (4.5.0)
Collecting nvidia-nvtx-cu11==11.7.91
  Downloading nvidia_nvtx_cu11-11.7.91-py3-none-manylinux1_x86_64.whl (98 kB)
                             98.6/98.6 KB
13.5 MB/s eta 0:00:00
Requirement already satisfied: sympy in /usr/local/lib/python3.9/dist-
packages (from torch) (1.11.1)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.9/dist-packages
(from torch) (3.1.2)
Requirement already satisfied: networkx in /usr/local/lib/python3.9/dist-
packages (from torch) (3.0)
Requirement already satisfied: filelock in /usr/local/lib/python3.9/dist-
packages (from torch) (3.10.2)
Collecting nvidia-cufft-cu11==10.9.0.58
  Downloading nvidia_cufft_cu11-10.9.0.58-py3-none-manylinux1_x86_64.whl (168.4
```

```
MB)
                              168.4/168.4
MB 7.6 MB/s eta 0:00:00
Collecting nvidia-cuda-runtime-cu11==11.7.99
  Downloading nvidia_cuda_runtime_cu11-11.7.99-py3-none-manylinux1_x86_64.whl
(849 kB)
                             849.3/849.3 KB
54.2 MB/s eta 0:00:00
Collecting nvidia-curand-cu11==10.2.10.91
  Downloading nvidia_curand_cu11-10.2.10.91-py3-none-manylinux1_x86_64.whl (54.6
MB)
                              54.6/54.6 MB
13.5 MB/s eta 0:00:00
Collecting nvidia-cuda-cupti-cu11==11.7.101
  Downloading nvidia_cuda_cupti_cu11-11.7.101-py3-none-manylinux1_x86_64.whl
(11.8 MB)
                              11.8/11.8 MB
88.2 MB/s eta 0:00:00
Collecting nvidia-cusolver-cu11==11.4.0.1
  Downloading nvidia_cusolver_cu11-11.4.0.1-2-py3-none-manylinux1_x86_64.whl
(102.6 MB)
                              102.6/102.6
MB 9.0 MB/s eta 0:00:00
Collecting triton==2.0.0
  Downloading
triton-2.0.0-1-cp39-cp39-manylinux2014_x86_64.manylinux_2_17_x86_64.whl (63.3
MB)
                              63.3/63.3 MB
11.3 MB/s eta 0:00:00
Collecting nvidia-nccl-cu11==2.14.3
  Downloading nvidia_nccl_cu11-2.14.3-py3-none-manylinux1_x86_64.whl (177.1 MB)
                              177.1/177.1
MB 7.9 MB/s eta 0:00:00
Collecting nvidia-cublas-cu11==11.10.3.66
  Downloading nvidia_cublas_cu11-11.10.3.66-py3-none-manylinux1_x86_64.whl
(317.1 MB)
                              317.1/317.1
MB 4.3 MB/s eta 0:00:00
Requirement already satisfied: setuptools in
/usr/local/lib/python3.9/dist-packages (from nvidia-cublas-
cu11==11.10.3.66->torch) (67.6.0)
Requirement already satisfied: wheel in /usr/local/lib/python3.9/dist-packages
(from nvidia-cublas-cu11==11.10.3.66->torch) (0.40.0)
Collecting lit
  Downloading lit-16.0.0.tar.gz (144 kB)
```

```
                                     145.0/145.0 KB
21.7 MB/s eta 0:00:00
  Preparing metadata (setup.py) … done
Requirement already satisfied: cmake in /usr/local/lib/python3.9/dist-packages
(from triton==2.0.0->torch) (3.25.2)
Requirement already satisfied: requests in /usr/local/lib/python3.9/dist-
packages (from torchdata) (2.27.1)
Requirement already satisfied: urllib3>=1.25 in /usr/local/lib/python3.9/dist-
packages (from torchdata) (1.26.15)
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.9/dist-
packages (from jinja2->torch) (2.1.2)
Requirement already satisfied: charset-normalizer~=2.0.0 in
/usr/local/lib/python3.9/dist-packages (from requests->torchdata) (2.0.12)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.9/dist-
packages (from requests->torchdata) (3.4)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.9/dist-packages (from requests->torchdata) (2022.12.7)
Requirement already satisfied: mpmath>=0.19 in /usr/local/lib/python3.9/dist-
packages (from sympy->torch) (1.3.0)
Building wheels for collected packages: lit
  Building wheel for lit (setup.py) … done
  Created wheel for lit: filename=lit-16.0.0-py3-none-any.whl size=93601
sha256=f58f34978957185f16ba02799e9144b203ec74f040cbab394fd93fb34c0a1808
  Stored in directory: /root/.cache/pip/wheels/c7/ee/80/1520ca86c3557f70e5504b80
2072f7fc3b0e2147f376b133ed
Successfully built lit
Installing collected packages: lit, nvidia-nvtx-cu11, nvidia-nccl-cu11, nvidia-
cusparse-cu11, nvidia-curand-cu11, nvidia-cufft-cu11, nvidia-cuda-runtime-cu11,
nvidia-cuda-nvrtc-cu11, nvidia-cuda-cupti-cu11, nvidia-cublas-cu11, nvidia-
cusolver-cu11, nvidia-cudnn-cu11, triton, torch, torchdata
ERROR: pip's dependency resolver does not currently take into account all
the packages that are installed. This behaviour is the source of the following
dependency conflicts.
torchvision 0.14.1+cu116 requires torch==1.13.1, but you have torch 2.0.0 which
is incompatible.
torchaudio 0.13.1+cu116 requires torch==1.13.1, but you have torch 2.0.0 which
is incompatible.
fastai 2.7.11 requires torch<1.14,>=1.7, but you have torch 2.0.0 which is
incompatible.
Successfully installed lit-16.0.0 nvidia-cublas-cu11-11.10.3.66 nvidia-cuda-
cupti-cu11-11.7.101 nvidia-cuda-nvrtc-cu11-11.7.99 nvidia-cuda-runtime-
cu11-11.7.99 nvidia-cudnn-cu11-8.5.0.96 nvidia-cufft-cu11-10.9.0.58 nvidia-
curand-cu11-10.2.10.91 nvidia-cusolver-cu11-11.4.0.1 nvidia-cusparse-
```

```
cu11-11.7.4.91 nvidia-nccl-cu11-2.14.3 nvidia-nvtx-cu11-11.7.91 torch-2.0.0
torchdata-0.6.0 triton-2.0.0
```

[ ]:
```
!pip3 install torchtext==0.6.0
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-
wheels/public/simple/
Collecting torchtext==0.6.0
  Downloading torchtext-0.6.0-py3-none-any.whl (64 kB)
                             64.2/64.2 KB
4.1 MB/s eta 0:00:00
Requirement already satisfied: numpy in /usr/local/lib/python3.9/dist-
packages (from torchtext==0.6.0) (1.22.4)
Requirement already satisfied: tqdm in /usr/local/lib/python3.9/dist-packages
(from torchtext==0.6.0) (4.65.0)
Collecting sentencepiece
  Downloading
sentencepiece-0.1.97-cp39-cp39-manylinux_2_17_x86_64.manylinux2014_x86_64.whl
(1.3 MB)
                             1.3/1.3 MB
28.7 MB/s eta 0:00:00
Requirement already satisfied: six in /usr/local/lib/python3.9/dist-
packages (from torchtext==0.6.0) (1.16.0)
Requirement already satisfied: requests in /usr/local/lib/python3.9/dist-
packages (from torchtext==0.6.0) (2.27.1)
Requirement already satisfied: torch in /usr/local/lib/python3.9/dist-packages
(from torchtext==0.6.0) (2.0.0)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.9/dist-packages (from requests->torchtext==0.6.0)
(2022.12.7)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in
/usr/local/lib/python3.9/dist-packages (from requests->torchtext==0.6.0)
(1.26.15)
Requirement already satisfied: charset-normalizer~=2.0.0 in
/usr/local/lib/python3.9/dist-packages (from requests->torchtext==0.6.0)
(2.0.12)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.9/dist-
packages (from requests->torchtext==0.6.0) (3.4)
Requirement already satisfied: nvidia-nccl-cu11==2.14.3 in
/usr/local/lib/python3.9/dist-packages (from torch->torchtext==0.6.0) (2.14.3)
Requirement already satisfied: networkx in /usr/local/lib/python3.9/dist-
packages (from torch->torchtext==0.6.0) (3.0)
Requirement already satisfied: nvidia-cuda-runtime-cu11==11.7.99 in
/usr/local/lib/python3.9/dist-packages (from torch->torchtext==0.6.0) (11.7.99)
Requirement already satisfied: nvidia-cusparse-cu11==11.7.4.91 in
/usr/local/lib/python3.9/dist-packages (from torch->torchtext==0.6.0)
(11.7.4.91)
Requirement already satisfied: nvidia-cublas-cu11==11.10.3.66 in
```

```
/usr/local/lib/python3.9/dist-packages (from torch->torchtext==0.6.0)
(11.10.3.66)
Requirement already satisfied: nvidia-cufft-cu11==10.9.0.58 in
/usr/local/lib/python3.9/dist-packages (from torch->torchtext==0.6.0)
(10.9.0.58)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.9/dist-packages
(from torch->torchtext==0.6.0) (3.1.2)
Requirement already satisfied: filelock in /usr/local/lib/python3.9/dist-
packages (from torch->torchtext==0.6.0) (3.10.2)
Requirement already satisfied: nvidia-cuda-cupti-cu11==11.7.101 in
/usr/local/lib/python3.9/dist-packages (from torch->torchtext==0.6.0) (11.7.101)
Requirement already satisfied: typing-extensions in
/usr/local/lib/python3.9/dist-packages (from torch->torchtext==0.6.0) (4.5.0)
Requirement already satisfied: sympy in /usr/local/lib/python3.9/dist-packages
(from torch->torchtext==0.6.0) (1.11.1)
Requirement already satisfied: nvidia-cusolver-cu11==11.4.0.1 in
/usr/local/lib/python3.9/dist-packages (from torch->torchtext==0.6.0) (11.4.0.1)
Requirement already satisfied: nvidia-nvtx-cu11==11.7.91 in
/usr/local/lib/python3.9/dist-packages (from torch->torchtext==0.6.0) (11.7.91)
Requirement already satisfied: nvidia-cudnn-cu11==8.5.0.96 in
/usr/local/lib/python3.9/dist-packages (from torch->torchtext==0.6.0) (8.5.0.96)
Requirement already satisfied: nvidia-cuda-nvrtc-cu11==11.7.99 in
/usr/local/lib/python3.9/dist-packages (from torch->torchtext==0.6.0) (11.7.99)
Requirement already satisfied: triton==2.0.0 in /usr/local/lib/python3.9/dist-
packages (from torch->torchtext==0.6.0) (2.0.0)
Requirement already satisfied: nvidia-curand-cu11==10.2.10.91 in
/usr/local/lib/python3.9/dist-packages (from torch->torchtext==0.6.0)
(10.2.10.91)
Requirement already satisfied: wheel in /usr/local/lib/python3.9/dist-packages
(from nvidia-cublas-cu11==11.10.3.66->torch->torchtext==0.6.0) (0.40.0)
Requirement already satisfied: setuptools in /usr/local/lib/python3.9/dist-
packages (from nvidia-cublas-cu11==11.10.3.66->torch->torchtext==0.6.0) (67.6.0)
Requirement already satisfied: lit in /usr/local/lib/python3.9/dist-packages
(from triton==2.0.0->torch->torchtext==0.6.0) (16.0.0)
Requirement already satisfied: cmake in /usr/local/lib/python3.9/dist-packages
(from triton==2.0.0->torch->torchtext==0.6.0) (3.25.2)
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.9/dist-
packages (from jinja2->torch->torchtext==0.6.0) (2.1.2)
Requirement already satisfied: mpmath>=0.19 in /usr/local/lib/python3.9/dist-
packages (from sympy->torch->torchtext==0.6.0) (1.3.0)
Installing collected packages: sentencepiece, torchtext
  Attempting uninstall: torchtext
    Found existing installation: torchtext 0.15.1
    Uninstalling torchtext-0.15.1:
      Successfully uninstalled torchtext-0.15.1
Successfully installed sentencepiece-0.1.97 torchtext-0.6.0
```

[11]:
```
import numpy as np
import pandas as pd
import spacy
from torch import nn
import torch
from torchtext import data
from torch.nn  import functional as F
import torch.optim as  optim
import seaborn as sb
import matplotlib.pyplot as plt
%matplotlib inline
```

[4]:
```
train = pd.read_csv('train.csv')
test = pd.read_csv('test.csv')
```

[5]:
```
print("Train shape:",train.shape)
print("Test shape:",test.shape)
```

```
Train shape: (31962, 3)
Test shape: (17197, 2)
```

[6]:
```
train.label.value_counts()
```

[6]:
```
0    29720
1     2242
Name: label, dtype: int64
```

[7]:
```
train.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
```

[8]:
```
train.head()
```

[8]:

	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

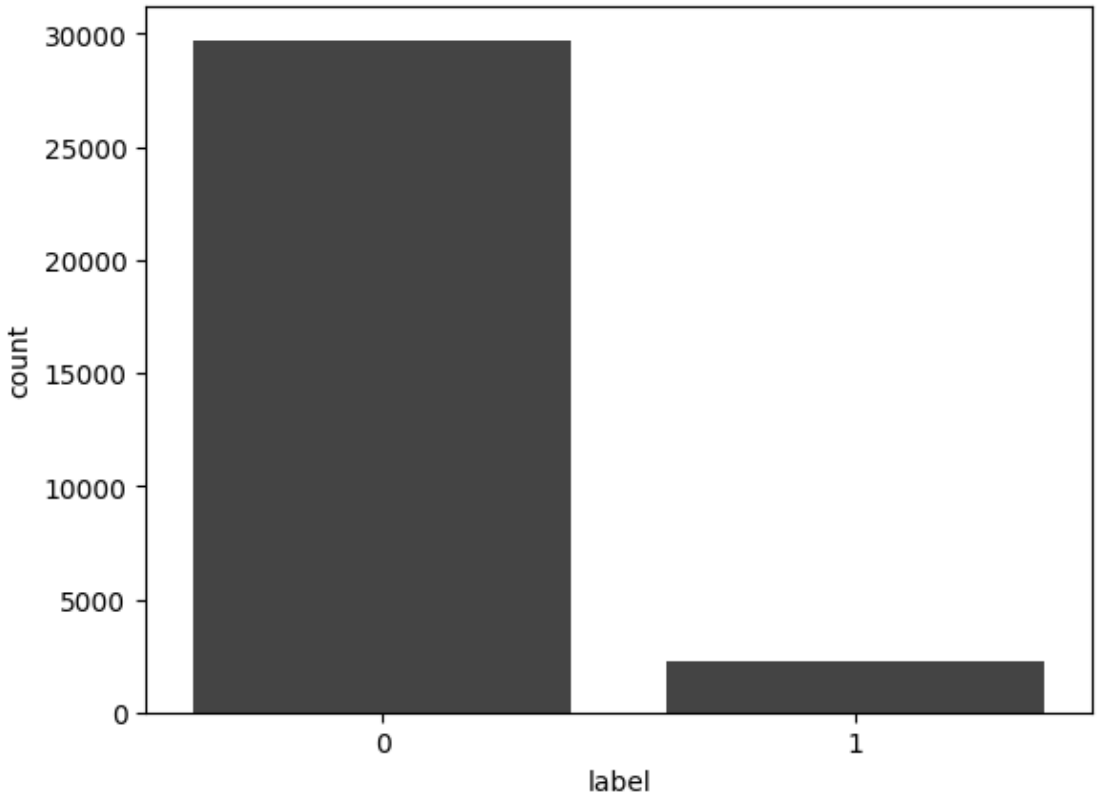
[9]:
```
test.head()
```

[9]:

	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…

1.2 EDA

[12]:
```
sb.countplot(data=train, x='label',color='#444444');
plt.show()
```



[28]:
```
hate_inices = train.loc[train['label']==1].index
```

[56]:
```
cleaned = train["tweet"].apply(myTokenizer)
clouds = cleaned.apply(lambda x: ' '.join(x))
```

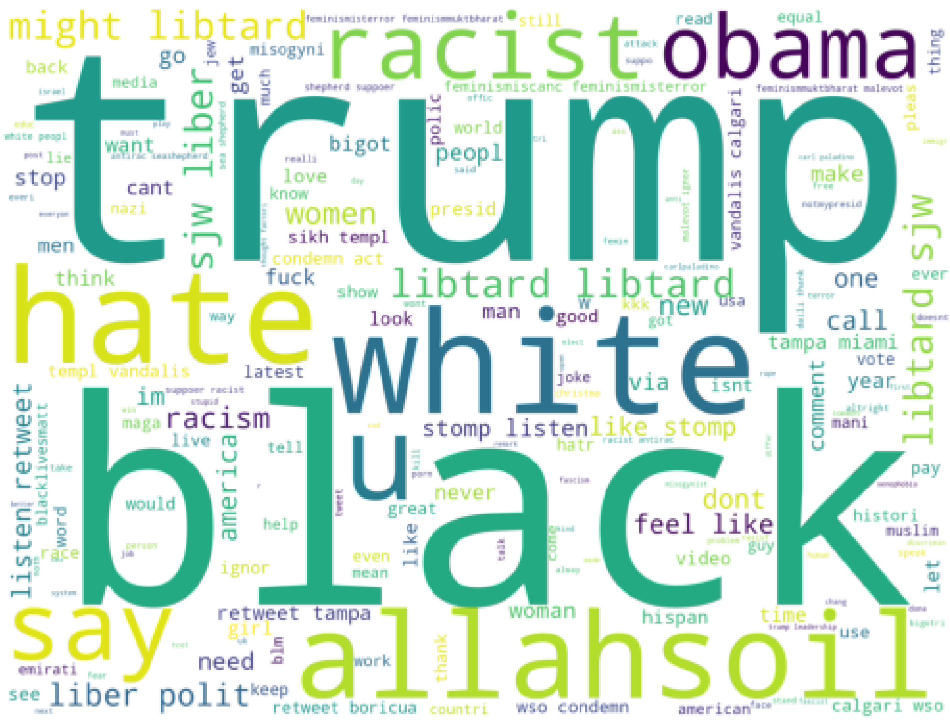
[57]:
```
hate_speech = clouds.loc[clouds.index.isin(hate_inices)]
```

[58]:
```
stops.append('user')
stops.append('users')
stops.append('amp')
```

[53]:
```
from wordcloud import WordCloud, STOPWORDS

text = " ".join(review for review in cleaned.astype(str))

wordcloud = WordCloud(width = 2000, height = 1500,
                background_color = 'white',
                stopwords = stops,
                min_font_size = 10).generate(text)
plt.imshow(wordcloud)
plt.axis("off")
plt.show()
```

[54]:
```
# HATE SPEECH

text = " ".join(review for review in hate_speech.astype(str))

wordcloud = WordCloud(width = 2000, height = 1500,
                background_color = 'white',
                stopwords = stops,
```

```
                  min_font_size = 10).generate(text)
plt.imshow(wordcloud)
plt.axis("off")
plt.show()
```




```
[55]: from gensim.models import Word2Vec
      # Build the model
      w2v_model = Word2Vec(min_count=5,
                           window=10,
                           vector_size = 1000,
                           sample=5e-5,
                           epochs = 15,
                           alpha=0.025,
                           min_alpha=0.0007,
                           negative=20)
```

```
[59]: w2v_model.build_vocab(cleaned.tolist(), progress_per=10000)
```

```
[62]: # You can use the generated files to view the embeddings using this url: https:/
       ↪/projector.tensorflow.org/
      savename = 'Speech'
      with open( savename + '_tensors.tsv',  'w+',encoding="utf-8") as tensors:
          with open ( savename + '_metadata.tsv', 'w+', encoding="utf-8") as metadata:
```

```
        for word in w2v_model.wv.index_to_key :
            encoded = word
            metadata.write(encoded + '\n')
            vector_row = '\t'.join(map(str, w2v_model.wv.get_vector(word)))
            tensors.write(vector_row + '\n')
```

[37]:
```
import nltk
nltk.download("punkt")

import re
from spacy.tokenizer import Tokenizer
from spacy.lang.en import English
nlp = English()

tokenizer = Tokenizer(nlp.vocab)

from nltk import word_tokenize,sent_tokenize
from nltk.stem  import PorterStemmer


from nltk.corpus import stopwords
from nltk.stem import SnowballStemmer

nltk.download('stopwords')
stops = stopwords.words("english")


def removepunc(my_str): # function to remove punctuation
    punctuations = '''!()-[]{};:'"\,<>./?@#$%^&*_~'''
    no_punct = ""
    for char in my_str:
        if char not in punctuations:
            no_punct = no_punct + char
    return no_punct

def hasNumbers(inputString):
    return bool(re.search(r'\d', inputString))
snowstem = SnowballStemmer("english")
portstem = PorterStemmer()
```

```
[nltk_data] Downloading package punkt to /root/nltk_data…
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to /root/nltk_data…
[nltk_data]   Package stopwords is already up-to-date!
```

The tokenizer we're implementing is a function that conducts basic text processing tasks such as converting the text to lowercase, eliminating punctuation, stop words, and numbers. Additionally, it eliminates unnecessary spaces and characters utilizing regex. Prior to utilizing the tokenizer, I

manually evaluated its performance on the training dataframe.

[17]:
```
def myTokenizer(x):
    return  [snowstem.stem(word.text)for word in
          tokenizer(removepunc(re.sub(r"\s+\s+"," ",re.sub(r"[^A-Za-z0-9()!?
 ↪@\'\`\"\r+\r+\n+\n+\b+]"," ",x.lower()))).strip())
          if (word.text not in stops and not hasNumbers(word.text)) ]
```

I am utilizing the torchtext fields and dataset classes to simplify the process of preparing the dataset for the PyTorch model. The most straightforward approach I've discovered for transforming a dataframe into a torchtext dataset is by utilizing the DataFrameDataset class. Please note that this particular cell may require some time to complete its execution.

[ ]:
```
TEXT = data.Field(tokenize=myTokenizer,batch_first=True,fix_length=140)
LABEL = data.LabelField(dtype=torch.float ,batch_first=True)


class DataFrameDataset(data.Dataset):

    def __init__(self, df, text_field, label_field, is_test=False, **kwargs):
        fields = [('comment_text', text_field), ('toxic', label_field)]
        examples = []
        for i, row in df.iterrows():
            if row['label'] == 0:
                label = 0
            else:
                label = 1

            text = row['tweet']
            examples.append(data.Example.fromlist([text, label], fields))

        super().__init__(examples, fields, **kwargs)


torchdataset = DataFrameDataset(train, TEXT,LABEL)
```

[ ]:
```
train_data, valid_data = torchdataset.split(split_ratio=0.8)

TEXT.build_vocab(train_data,min_freq=3)
LABEL.build_vocab(train_data)

#No. of unique tokens in text
print("Size of TEXT vocabulary:",len(TEXT.vocab))

#No. of unique tokens in label
print("Size of LABEL vocabulary:",len(LABEL.vocab))

#Commonly used words
```

```
print(TEXT.vocab.freqs.most_common(10))

print(len(TEXT.vocab))
```

```
Size of TEXT vocabulary: 6810
Size of LABEL vocabulary: 2
[('user', 14147), ('love', 2619), ('day', 2302), (' ', 1867), ('happi', 1694),
('amp', 1421), ('thank', 1252), ('time', 1021), ('get', 1002), ('u', 973)]
6810
```

[ ]: