```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
# from wordcloud import WordCloud

import nltk
nltk.download("punkt")
nltk.download("wordnet")
nltk.download("stopwords")

from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer

from sklearn.model_selection import train_test_split
from tensorflow.keras.preprocessing.text import Tokenizer
from tensorflow.keras.preprocessing import sequence #unique id

from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense, SimpleRNN, Dropout, Embedding
import warnings
warnings.filterwarnings("ignore")
```

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

Automatic saving failed. This file was updated remotely or in another tab. Show diff

```
df.head()
```

| | Index | message to examine | label (depression result) |
|---|---|---|---|
| **0** | 106 | just had a real good moment. i missssssssss hi... | 0 |
| **1** | 217 | is reading manga http://plurk.com/p/mzp1e | 0 |
| **2** | 220 | @comeagainjen http://twitpic.com/2y2lx - http:... | 0 |
| **3** | 288 | @lapcat Need to send 'em to my accountant tomo... | 0 |
| **4** | 540 | ADD ME ON MYSPACE!!! myspace.com/LookThunder | 0 |

```
df.isnull().sum()
```

```
    Index                        0
    message to examine           0
    label (depression result)    0
    dtype: int64
```

```
df = df.drop(['Index'],axis=1)
```

```
df['label (depression result)'].value_counts()
```

```
    0    8000
    1    2314
    Name: label (depression result), dtype: int64
```

```
def cleantext(text):
  tokens = word_tokenize(text.lower())
  ftoken = [t for t in tokens if(t.isalpha())]
  stop = stopwords.words("english")
  ctoken = [t for t in ftoken if(t not in stop)]
  lemma = WordNetLemmatizer()
  ltoken = [lemma.lemmatize(t) for t in ctoken]
  return " ".join(ltoken)
```

```
df['message to examine']=df['message to examine'].apply(cleantext)
```

```
sentlen = []

for sent in df["message to examine"]:
  sentlen.append(len(word_tokenize(sent)))

df["SentLen"] = sentlen
df.head()
```

| | message to examine | label (depression result) | SentLen |
|---|---|---|---|
| **0** | real good moment miss much | 0 | 5 |
| **1** | reading manga http | 0 | 3 |
| **2** | comeagainjen http http | 0 | 3 |
| **3** | lapcat need send accountant tomorrow oddly eve... | 0 | 12 |
| **4** | add myspace | 0 | 2 |

```
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
x = df['message to examine']
y = le.fit_transform(df['label (depression result)'])
```

```
np.quantile(sentlen, 0.95)

     19.0
```

Automatic saving failed. This file was updated remotely or in another tab. Show diff

```
max_len = np.quantile(sentlen, 0.95)
```

```
max(df['SentLen'])

     57
```

```
xtrain,xtest,ytrain,ytest = train_test_split(x,y,test_size=0.30,random_state=1)
```

```
tok = Tokenizer(char_level=False, split=" ")
tok.fit_on_texts(xtrain)
```

```
vocab_len = len(tok.index_word)
vocab_len

     13499
```

```
seqtrain = tok.texts_to_sequences(xtrain) #step1
seqtrain
```

```
      [34, 191, 3, 5687, 103],
      [3342, 110],
      [3343,
       54,
       68,
       5688,
       134,
       45,
       428,
       942,
       2520,
       5689,
       6,
       3344,
       102,
       40,
       3343,
       10,
       390,
       608,
       16,
       1,
       186,
       527,
       359,
       517,
       2024,
       5690,
       54,
       8],
      [5691, 36, 6, 687],
      [36, 1465, 111, 2025, 52, 133, 1415, 3345, 216],
      [679, 1466, 1716, 5692, 102],
      [92, 1717, 13, 319],
      [5693, 172, 5694, 1142, 327, 5695, 3346, 1143, 5696, 31, 2521, 5697, 37],
```

Automatic saving failed. This file was updated remotely or in another tab. Show diff

```
      [5700, 168, 9, 200, 1114, 5701, 67, 200, 3347, 368, 320, 1298],
      ...]
```

```
seqmattrain = sequence.pad_sequences(seqtrain, maxlen= int(max_len)) #step2
seqmattrain
```

```
    array([[    0,     0,     0, ...,    32,     4,  4492],
           [    0,     0,     0, ...,    31,   744,  1053],
           [    0,     0,     0, ...,    24,   105,  4493],
           ...,
           [    0,     0,     0, ...,    36,    88,   173],
           [    0,     0,     0, ...,    11, 13497,   131],
           [    0,     0,     0, ...,    29, 13499,   934]], dtype=int32)
```

```
seqtest = tok.texts_to_sequences(xtest)
seqmattest = sequence.pad_sequences(seqtest, maxlen=int(max_len))
```

```
from imblearn.over_sampling import SMOTE
sm = SMOTE(sampling_strategy='minority',random_state=34)
xsmaple,ysample = sm.fit_resample(seqmattrain,ytrain)
```

```
pd.DataFrame({'ysample':ysample}).value_counts()
```

```
    ysample
    0          5583
    1          5583
    dtype: int64
```

```
rnn = Sequential()

rnn.add(Embedding(vocab_len+1,50, input_length=int(max_len), mask_zero=True))
rnn.add(SimpleRNN(units=32, activation="tanh"))
rnn.add(Dense(units=32, activation="relu"))
rnn.add(Dropout(0.2))

rnn.add(Dense(units=1, activation="sigmoid"))
rnn.compile(optimizer="adam", loss='binary_crossentropy')

rnn.fit(xsmaple, ysample, batch_size=50, epochs=25)

ypred = rnn.predict(seqmattest)
```

```
    Epoch 1/25
    224/224 [==============================] - 8s 24ms/step - loss: 0.4522
    Epoch 2/25
    224/224 [==============================] - 4s 19ms/step - loss: 0.1947
    Epoch 3/25
    224/224 [==============================] - 4s 19ms/step - loss: 0.0619
    Epoch 4/25
    224/224 [==============================] - 5s 24ms/step - loss: 0.0215
    Epoch 5/25
    224/224 [==============================] - 4s 19ms/step - loss: 0.0113
    Epoch 6/25
    224/224 [==============================] - 5s 20ms/step - loss: 0.0050
    Epoch 7/25
    224/224 [==============================] - 5s 22ms/step - loss: 0.0033
    Epoch 8/25
    224/224 [==============================] - 4s 19ms/step - loss: 0.0026
    Epoch 9/25
    224/224 [==============================] - 5s 22ms/step - loss: 0.0024
    Epoch 10/25
    224/224 [==============================] - 5s 21ms/step - loss: 0.0022
    Epoch 11/25
    224/224 [==============================] - 4s 19ms/step - loss: 0.0020
    Epoch 12/25
    224/224 [==============================] - 5s 24ms/step - loss: 0.0019
    Epoch 13/25
    224/224 [==============================] - 4s 19ms/step - loss: 0.0019
```

Automatic saving failed. This file was updated remotely or in another tab. Show diff

```
    Epoch 15/25
    224/224 [==============================] - 5s 24ms/step - loss: 0.0017
    Epoch 16/25
    224/224 [==============================] - 4s 19ms/step - loss: 0.0019
    Epoch 17/25
    224/224 [==============================] - 4s 19ms/step - loss: 0.0018
    Epoch 18/25
    224/224 [==============================] - 5s 24ms/step - loss: 0.0018
    Epoch 19/25
    224/224 [==============================] - 4s 19ms/step - loss: 0.0016
    Epoch 20/25
    224/224 [==============================] - 4s 19ms/step - loss: 0.0017
    Epoch 21/25
    224/224 [==============================] - 5s 24ms/step - loss: 0.0017
    Epoch 22/25
    224/224 [==============================] - 4s 19ms/step - loss: 0.0017
    Epoch 23/25
    224/224 [==============================] - 5s 24ms/step - loss: 0.0016
    Epoch 24/25
    224/224 [==============================] - 4s 20ms/step - loss: 0.0017
    Epoch 25/25
    224/224 [==============================] - 4s 19ms/step - loss: 0.0016
    97/97 [==============================] - 1s 3ms/step
```

```
ypred = np.where(ypred>0.5,1,0)
```

```
from sklearn.metrics import classification_report
print(classification_report(ytest,ypred))
```

```
                   precision    recall  f1-score   support

               0       1.00      0.82      0.90      2417
               1       0.61      0.99      0.75       678

        accuracy                           0.86      3095
       macro avg       0.80      0.90      0.83      3095
    weighted avg       0.91      0.86      0.87      3095
```

✓ 0s completed at 6:13 PM ● ×

Automatic saving failed. This file was updated remotely or in another tab. Show diff