!pip install transformers

```
Requirement already satisfied: transformers in /usr/local/lib/python3.6/dist-package
     Requirement already satisfied: packaging in /usr/local/lib/python3.6/dist-packages (
     Requirement already satisfied: requests in /usr/local/lib/python3.6/dist-packages (f
     Requirement already satisfied: dataclasses; python_version < "3.7" in /usr/local/lib
     Requirement already satisfied: numpy in /usr/local/lib/python3.6/dist-packages (from
     Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.6/dist-packages
     Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.6/dist-pa
     Requirement already satisfied: protobuf in /usr/local/lib/python3.6/dist-packages (f
     Requirement already satisfied: tokenizers==0.9.2 in /usr/local/lib/python3.6/dist-pa
     Requirement already satisfied: sentencepiece!=0.1.92 in /usr/local/lib/python3.6/dis
     Requirement already satisfied: sacremoses in /usr/local/lib/python3.6/dist-packages
     Requirement already satisfied: filelock in /usr/local/lib/python3.6/dist-packages (f
     Requirement already satisfied: six in /usr/local/lib/python3.6/dist-packages (from p
     Requirement already satisfied: pyparsing>=2.0.2 in /usr/local/lib/python3.6/dist-pac
     Requirement already satisfied: chardet<4,>=3.0.2 in /usr/local/lib/python3.6/dist-pa
     Requirement already satisfied: idna<3,>=2.5 in /usr/local/lib/python3.6/dist-package
     Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in /usr/local
     Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.6/dist-p
     Requirement already satisfied: setuptools in /usr/local/lib/python3.6/dist-packages
     Requirement already satisfied: click in /usr/local/lib/python3.6/dist-packages (from
     Requirement already satisfied: joblib in /usr/local/lib/python3.6/dist-packages (fro
import pandas as pd
import numpy as np
from tqdm import tqdm
from tqdm.notebook import tqdm_notebook
tqdm notebook.pandas()
import warnings
warnings.filterwarnings('ignore')
import tensorflow as tf
from tqdm import tqdm
from google.colab import drive
drive.mount('/content/drive')
     Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.
#! cp '/content/drive/My Drive/tweet-sentiment-extraction/preprocessed train.csv'
#! cp '/content/drive/My Drive/tweet-sentiment-extraction/preprocessed test.csv'
! cp '/content/drive/My Drive/tweet-sentiment-extraction/train.csv'
! cp '/content/drive/My Drive/tweet-sentiment-extraction/test.csv'
#train df = pd.read csv('preprocessed train.csv')
#test_df = pd.read_csv('preprocessed_test.csv')
train_df = pd.read_csv('train.csv')
test df = pd.read csv('test.csv')
train_df.shape,test_df.shape
```

```
((27481, 4), (3534, 3))
```

train_df.sample(10)

sentiment	selected_text	text	textID	
positive	Wow! I am so proud Its great.	I AM OFFICIALLY CELEBRATING MY 400TH TWEET. Wo	8f3ce0745c	23099
negative	broken	my computers broken and I broke my laptop so	653eaec42c	18082
neutral	It`s cold and cloudy and I have a million load	It`s cold and cloudy and I have a million load	d05cc52d1a	8563
positive	interesting.	back in Durango. Chaperoning a party. interest	ddc208fbff	11727
positive	welcome to dk #jaoo	and welcome to dk #jaoo	e5b3f605e0	18752
neutral	No news yet	No news yet	0da839bf7c	21307
neutral	Im am going to send some twits to some folk	Im am going to send some twits to some folk	a3df1e8e30	15246

test_df.sample(5)

sentiment	text	textID	
negative	Musem equaled FAIL!!!!! hall of **** life was	846d43bc8f	1896
neutral	need some Pizzacan`t get out of office	9630a8a60e	2399
positive	did you tweet me? i thought i remembered seei	93b5d1bbea	967
negative	@_supernatural_ http://twitpic.com/66l83 - I r	f6f7382036	1485
positive	sounds like Jeff has the best job in the	5d47d76ed3	3061

```
train_df.isnull().sum(),test_df.isnull().sum()
```

```
(textID
                  0
text
                  1
selected_text
                  1
sentiment
dtype: int64, textID
text
sentiment
dtype: int64)
```

train_df[train_df.text.isna()]

	textID	text	selected_text	sentiment
314	fdb77c3752	NaN	NaN	neutral

```
train_df.dropna(axis=0,inplace=True)
train df.shape, test df.shape
```

```
((27480, 4), (3534, 3))
```

train_df.sample(7)

sentiment	selected_text	text	textID	
neutral	that`s such a long time from now	that`s such a long time from now	567b068131	10610
positive	Haha	Yes! Haha and then when he impaled himself wi	a34a224f03	15730
positive	Enjoy	Going to be buckling down this week. So, relie	17aeaef146	17009
negative	sad	Why I gotta be all that? I sad now	73355afc38	4349
:4:	المحمل المنطرة	good luck chan! gue kmrn	4000F-00-F	00400

train_df['text'] = train_df['text'].progress_apply(lambda x : str(x).lower())
train_df['selected_text'] = train_df['selected_text'].progress_apply(lambda x : str(x).low

100% 27480/27480 [00:00<00:00, 297789.62it/s]

100% 27480/27480 [00:00<00:00, 279547.70it/s]

#https://www.kaggle.com/cdeotte/tensorflow-roberta-0-705
#Reference notebook for Tokenizing ideas

train_df.sample(10)

```
textID
                                              text
                                                                 selected_text sentiment
                            hig hair rock day today me
                                                        hin hair rock day today me
train_df[train_df.text=='']
        textID text selected_text sentiment
                            ooo growl and vummy time
                                                        ooo growl and vummy time
             ----
from sklearn.model_selection import train_test_split
x_train, x_val , y_train , y_val = train_test_split(train_df[['text','sentiment']],train_c
x_train.shape, x_val.shape , y_train.shape , y_val.shape
     ((21984, 2), (5496, 2), (21984,), (5496,))
from transformers import RobertaTokenizer
tokenizer = RobertaTokenizer.from_pretrained('roberta-base',add_prefix_space=True)
tokenizer.encode(' hello world'),tokenizer.encode('hello world')
     ([0, 20760, 232, 2], [0, 20760, 232, 2])
MAX LEN=128
count = x_train.shape[0]
input_ids = np.zeros((count,MAX_LEN),dtype='int32')
attention_mask = np.zeros((count,MAX_LEN),dtype='int32')
start_tokens = np.zeros((count,MAX_LEN),dtype='int32')
end_tokens = np.zeros((count,MAX_LEN),dtype='int32')
#Understanding tokenizer encode
print(x_train['text'].values[10])
print(x_train['sentiment'].values[10])
v = tokenizer.encode_plus(x_train['text'].values[10],x_train['sentiment'].values[10],add_s
                          pad_to_max_length=True,return_tensors='tf',verbose=False)
print(v['input ids'])
print(v['attention_mask'])
```

```
awww the boys have no beards?
                                      i liked their manly man-ness [via ] #asylm
     positive
from tqdm import tqdm
for i,each in tqdm(enumerate(x_train.values)):
 val = tokenizer.encode_plus(each[0],each[1],add_special_tokens=True,max_length=128,retur
  input_ids[i] = val['input_ids']
  attention_mask[i] = val['attention_mask']
     21984it [00:08, 2496.82it/s]
                 1
                       1
                                   1
                                         1
                                                1
                                                      1
                                                            1
                                                                        1
                                                                              1
     tf.Tensor(
# Understanding the concepts of offset and creating start and end tokens
import random
st_tok = np.zeros((1,MAX_LEN),dtype='int32')
end_tok = np.zeros((1,MAX_LEN),dtype='int32')
k = random.randint(0,x_train.shape[0])
text1 = " "+" ".join(x_train['text'].values[k].split())
text2 = " ".join(y_train.values[k].split())
print(text1)
print(text2)
idx = text1.find(text2)
print(idx)
chars = np.zeros((len(text1)))
chars[idx:idx+len(text2)]=1
if text1[idx-1]==' ':
 chars[idx-1] = 1
print(chars,len(chars))
enc = tokenizer.encode(text1,add_special_tokens=False)
print(enc)
offsets = []; idx=0
tokens=[]
for t in enc:
 w = tokenizer.decode([t])
 tokens.append(w)
 offsets.append((idx,idx+len(w)))
 idx += len(w)
print(tokens)
print(offsets)
toks = []
for i,(a,b) in enumerate(offsets):
 sm = np.sum(chars[a:b])
 if sm>0:
   toks.append(i)
print(toks)
if len(toks)>0:
  st tok[0,toks[0]+1] = 1
```

```
end_{tok}[0, toks[-1]+1] = 1
print('Start tokens',st tok)
print('End tokens ',end_tok)
a = np.argmax(st tok)
b = np.argmax(end_tok)
print(a,b)
text1 = " "+" ".join(x_train['text'].values[k].split())
enc = tokenizer.encode(text1)
st = tokenizer.decode(enc[a:b+1])
print(st)
   star trek was great yet minor details needed to be worked out, but i give it a thum
   great
   15
   [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 1. 1. 1. 1. 1. 0. 0. 0. 0.
   0. 0. 0. 0. 0. 0. 102
   [999, 18526, 21, 372, 648, 3694, 1254, 956, 7, 28, 1006, 66, 6, 53, 939, 492, 24, 10
   [' star', ' trek', ' was', ' great', ' yet', ' minor', ' details', ' needed', ' to',
   [(0, 5), (5, 10), (10, 14), (14, 20), (20, 24), (24, 30), (30, 38), (38, 45), (45, 4)]
   [3]
   4 4
   great
from tqdm import tqdm
for k in tqdm(range(x_train.shape[0])):
 text1 = " "+" ".join(x_train['text'].values[k].split())
 text2 = " ".join(y_train.values[k].split())
 idx = text1.find(text2)
 chars = np.zeros((len(text1)))
 chars[idx:idx+len(text2)]=1
 if text1[idx-1]==' ':
  chars[idx-1] = 1
 enc = tokenizer.encode(text1,add special tokens=False)
 offsets = []; idx=0
 for t in enc:
  w = tokenizer.decode([t])
  offsets.append((idx,idx+len(w)))
  idx += len(w)
```

```
toks = []
 for i,(a,b) in enumerate(offsets):
   sm = np.sum(chars[a:b])
    if sm>0:
      toks.append(i)
 if len(toks)>0:
   start_tokens[k,toks[0]+1] = 1
   end_tokens[k,toks[-1]+1] = 1
     100%| 21984/21984 [00:09<00:00, 2271.47it/s]
#Visualize the results
import random
for _ in range(35,50):
  i = random.randint(0,x_train.shape[0])
  print(x_train.iloc[i]['text'])
 print(y_train.values[i])
  print([tokenizer.decode([x]) for x in tokenizer.encode(x_train.iloc[i]['text'])])
  print('Input ids',input_ids[i])
  print('attention mask',attention_mask[i])
  print('Start tokens',start_tokens[i])
 print('end Tokens',end_tokens[i])
  a = np.argmax(start_tokens[i])
 b=np.argmax(end_tokens[i])
  print(a,b)
 text1 = " "+" ".join(x_train.iloc[i]['text'].split())
  enc = tokenizer.encode(text1)
 st = tokenizer.decode(enc[a:b+1])
  print(st)
 print('*'*50)
```

```
unfortunately, that's how it usually goes. in answer to yesterday...i'm okay. thin
unfortunately,
           ',', ' that', '`', 's', ' how', ' it', ' usually', ' goes'
['<s>', ' unfortunately',
      0 9574
             14 12905
                  29 141
                       24 2333 1411
Input ids [
           6
       7 2350
             118 12905
                     8578
                          383
 11 1948
           734
                   119
                        4
                             189
 28
   546
      62
         4
          2649
             1686
                 7
                   47
                      4
                         2
                           2
                            2430
  2
    1
       1
         1
            1
              1
                 1
                   1
                      1
                         1
                           1
                              1
  1
       1
    1
         1
            1
              1
                 1
                   1
                      1
                         1
                           1
                              1
  1
       1
         1
              1
                 1
                   1
                      1
                              1
  1
       1
              1
                 1
                      1
                              1
    1
         1
            1
                   1
                         1
                           1
  1
    1
       1
         1
            1
              1
                 1
                   1
                      1
                         1
                           1
                              1
  1
       1
         1
              1
                 1
                      1
                        1
                           1
                              1
    1
            1
                   1
         1
              1
                           1
  1
       1
            1
                   1
  1
    1
         1
              1
                 1
                   1]
       1
            1
unfortunately,
***************
have $#@&!!! malware on my work pc! i miss having an it dept
i miss having an it dept
['<s>', ' have', ' $', '#', '@', '&', '!!!', ' malware', ' on', ' my', ' work', ' pc
Input ids [
        33
          68 10431 1039
                  947 16506 16886
                          15
                            127
                              173 46213
      939
        2649
                            2430
 328 1437
           519
              41
                   263
                     3320
                        2
                           2
                24
  2
    1
       1
         1
            1
              1
                 1
                   1
                      1
                           1
                              1
  1
    1
       1
         1
              1
                 1
                      1
                         1
                              1
            1
                   1
                           1
         1
                 1
    1
       1
            1
              1
                   1
                      1
                         1
                           1
                              1
  1
    1
       1
         1
            1
              1
                 1
                   1
                      1
                        1
                           1
  1
       1
         1
                 1
                      1
                              1
    1
            1
              1
                   1
                        1
                           1
  1
    1
       1
         1
            1
              1
                 1
                   1
                      1
                         1
                           1
                              1
  1
    1
       1
         1
            1
              1
                 1
                   1
                      1
                         1
                           1
                              1
  1
    1
       1
         1
            1
              1
                 1
                   1
                      1
                         1
                           1
                              1
    1
       1
         1
            1
              1
                 1
                   1]
13 19
i miss having an it dept
hm... i don't i can recommend any white chocolates though.. you have to move to the
```

hm... i don't i can recommend any white chocolates though.. you have to move to the ['<s>', ' h', 'm', '...', ' i', ' don', '`', 't', ' i', ' can', ' recommend', ' any'

```
119
               734
                  939
                     218 12905
                             939
                                   5940
                                      143
Input ids [
       0 1368
                                 64
               7586
                            517
1104 1855 23558
                   47
                      33
                          7
                               7
         1626
             600
                                  5
22209 30423 12905
          526
             78
               7586
                    2
                       2
                        7974
                             2
                               1
                                  1
     1
        1
           1
                                  1
  1
              1
                 1
                    1
                       1
                          1
                             1
                                1
  1
     1
        1
           1
              1
                 1
                    1
                       1
                          1
                             1
                               1
                                  1
  1
        1
           1
              1
                 1
                    1
                       1
                          1
                                  1
  1
        1
           1
                 1
                    1
                                  1
     1
              1
                       1
                          1
                             1
                               1
  1
           1
              1
                 1
                    1
                          1
                             1
                                1
                                  1
     1
        1
                       1
                          1
                             1
                               1
                                  1
  1
     1
        1
           1
              1
                 1
                    1
                       1
           1
                                1
        1
              1
                 1
                       1
                                  1
  1
     1
           1
                 1
                       1]
        1
              1
                    1
1 29
hm... i don't i can recommend any white chocolates though.. you have to move to the
              mcfly parody by chris moyles ****
http://tinyurl.com/d53dmn
http://tinyurl.com/d53dmn
              mcfly parody by chris moyles
['<s>', 'http', '://', 'tiny', 'url', '.', 'com', '/', 'd', '53', 'd', 'mn', ' ',
       0 2054
                           73
Input ids [
            640 40743 6423
                      4
                        175
                             417
                               4540
                                   417 17673
      1437
         1437 44355 19252 32806
                        1855
                           4663
                              475
1437 1437
                      30
                                 2160
 1634 31095
        2
           2
            7974
                 2
                          1
                       1
                                1
  1
     1
        1
           1
              1
                 1
                    1
                          1
                             1
                                1
                                  1
                       1
           1
                    1
  1
     1
        1
              1
                 1
                       1
                          1
                             1
                                1
                                  1
  1
     1
        1
           1
              1
                 1
                    1
                       1
                          1
                             1
                               1
  1
     1
        1
           1
              1
                 1
                    1
                       1
                          1
                             1
                               1
                                  1
  1
           1
                                  1
     1
        1
              1
                 1
                    1
                       1
                          1
                             1
                               1
  1
     1
        1
           1
              1
                 1
                    1
                       1
                          1
                             1
                               1
                                  1
  1
     1
        1
           1
              1
                 1
                    1
                       1
                          1
                             1
                                1
                                  1
           1
              1
                 1
                    1
                       1]
1 20
http://tinyurl.com/d53dmn mcfly parody by chris moyles
i can`t feel the weekend fever anymore when everyday is a day off
i can`t feel the weekend fever anymore when everyday is a day off
['<s>', ' i', ' can', '`', 't', ' feel', ' the', ' weekend', ' fever', ' anymore', ' Input ids [ 0 939 64 12905 90 619 5 983 11696 5988 77 7476
 16 1437
       10
          183
             160
                 2
                    2
                     2430
                          2
                             1
                               1
                                  1
  1
        1
           1
              1
                 1
                    1
                       1
                          1
                                1
                                  1
           1
              1
                 1
                    1
                       1
```

```
1
    1
      1
         1
           1
              1
                1
                   1
                     1
                        1
                          1
                             1
  1
    1
      1
         1
           1
              1
                1
                   1
                     1
                        1
                          1
                             1
  1
    1
      1
         1
           1
              1
                1
                     1
                        1
                          1
                             1
                   1
  1
    1
      1
         1
           1
              1
                1
                   1
                     1
                        1
                          1
                             1
                             1
  1
    1
      1
         1
           1
              1
                1
                   1
                     1
                        1
                          1
         1
              1
                     1
                          1
                             1
  1
      1
           1
                1
                   1
  1
    1
      1
         1
           1
              1
                1
                   1]
end
1 15
i can't feel the weekend fever anymore when everyday is a day off
maybe used to have. besides without ac it's too hot to sleep
['<s>', ' maybe', ' used', ' to', ' have', '.', ' besides', ' without', ' ac',
             7
      0 2085
                  4 12035
Input ids [
          341
               33
                      396 4285
                           24 12905
 350
        3581
               2430
                             1
  2131
      7
           2
              2
                   2
                     1
                        1
                          1
  1
      1
         1
              1
                1
                   1
                     1
                        1
                          1
                             1
  1
    1
      1
         1
              1
                1
                     1
                        1
                             1
           1
                   1
                          1
  1
      1
         1
           1
              1
                1
                     1
                        1
                          1
                             1
    1
                   1
  1
      1
         1
              1
                1
                        1
                          1
                             1
    1
           1
                   1
                     1
  1
         1
                             1
    1
      1
           1
              1
                1
                   1
                     1
                        1
                          1
                             1
  1
    1
      1
         1
           1
              1
                1
                     1
                        1
                          1
                   1
  1
    1
      1
         1
           1
              1
                1
                   1
                     1
                        1
                          1
                             1
  1
         1
                     1
                        1
                          1
                             1
    1
      1
           1
              1
                1
                   1
  1
      1
         1
                1
                   1]
    1
           1
              1
12 13
too hot
   ************
i don't understand anything, help
help
['<s>', ' i', ' don', '`', 't', ' understand', ' anything', ',', ' help', '</s>']
      0
       939
          218 12905
               90 1346
                    932
                        244
                            2
                              2
Input ids [
                       6
                               2430
  2
    1
      1
         1
           1
              1
                1
                   1
                     1
                        1
                          1
                             1
  1
    1
      1
         1
           1
              1
                1
                   1
                     1
                        1
                          1
                             1
  1
         1
                1
                     1
                             1
    1
      1
           1
              1
                   1
                        1
                          1
  1
    1
       1
         1
           1
              1
                1
                   1
                     1
                        1
                          1
                             1
                             1
  1
    1
      1
         1
              1
                1
                     1
                          1
           1
                   1
                        1
      1
         1
                             1
```

```
1
    1
       1
          1
               1
                  1
                    1
                       1
                             1
                               1
  1
       1
          1
            1
               1
                  1
                    1
                       1
                          1
                             1
                               1
    1
  1
    1
       1
          1
            1
               1
                  1
                    1
                       1
                          1
                             1
                               1
    1
       1
          1
               1
                  1
                    1]
            1
end
8 8
help
**************
can i bring yself to go for the 1000 heroism badges achievement?
can i bring yself to go for the 1000 heroism badges achievement?
['<s>', ' can', ' i', ' bring', ' y', 'self', ' to', ' go', ' for', ' the', ' 1000',
           939
             836 1423 13367
Input ids [
      0
        64
                        213
                           13
                              5 10775 38047
33592 8312
          2
             7974
                  2
                          1
                               1
      116
            2
                    1
                       1
                             1
  1
    1
       1
          1
            1
               1
                  1
                     1
                       1
                          1
                             1
                               1
  1
    1
       1
          1
            1
               1
                  1
                       1
                          1
                               1
                    1
                             1
  1
    1
       1
          1
            1
               1
                  1
                     1
                          1
                             1
                               1
  1
       1
               1
                  1
                          1
                             1
                               1
    1
          1
            1
                    1
                       1
  1
          1
               1
                       1
                               1
    1
       1
            1
                  1
                    1
                             1
                               1
  1
    1
       1
          1
            1
               1
                  1
                       1
                          1
                             1
                    1
  1
    1
       1
          1
            1
               1
                  1
                    1
                       1
                          1
                             1
                               1
  1
       1
          1
                       1
                          1
                             1
                               1
    1
            1
               1
                  1
                    1
    1
       1
          1
            1
               1
                  1
                    1]
1 14
can i bring yself to go for the 1000 heroism badges achievement?
ugh, you people and your fancy trips to italy while i`m slaving away at this compu
ugh
['<s>', ' ', ' u', 'gh', ',', ' you', ' people', ' and', ' your', ' fancy', ' trips'
Input ids [
      0 1437 1717 4147
                 6
                   47
                         8
                           110 13185 6734
                      82
         939 12905
                3369 13286
 24 7776
      150
              119
                      409
                         23
                            42
                              3034
  4
       2
        2430
    2
            2
               1
                  1
                    1
                       1
                          1
                             1
                               1
  1
       1
          1
               1
                  1
                             1
                               1
    1
            1
                    1
                       1
                          1
  1
    1
       1
          1
            1
               1
                  1
                     1
                       1
                          1
                             1
                               1
  1
    1
       1
          1
            1
               1
                  1
                    1
                       1
                          1
                             1
                               1
  1
    1
       1
          1
            1
               1
                  1
                    1
                       1
                          1
                             1
                               1
  1
    1
       1
          1
               1
                  1
                       1
                          1
                               1
            1
                     1
                             1
  1
    1
          1
            1
               1
                  1
                       1
                          1
                             1
                               1
       1
                     1
  1
    1
       1
          1
            1
               1
                  1
                       1
                          1
                             1
                               1
                     1
       1
          1
            1
               1
                     11
```

```
ugh
***************
sneezing frequently
sneezing frequently
['<s>', ' sne', 'ez', 'ing', ' frequently', '</s>']
    0 18013 5841 154 5705
                2 7974
              2
                    2
Input ids [
                      1
                          1
     1
       1
         1
             1
 1
   1
           1
               1
                 1
                  1
                    1
                      1
             1
 1
   1
     1
       1
         1
           1
               1
                 1
                   1
                    1
                      1
 1
   1
     1
       1
         1
           1
             1
               1
                 1
                   1
                      1
                    1
 1
   1
     1
       1
         1
           1
             1
               1
                   1
                    1
 1
   1
     1
       1
         1
           1
             1
                 1
                  1
                    1
                      1
               1
 1
     1
       1
         1
           1
                 1
               1
 1
     1
       1
           1
                 1
                      1
   1
         1
             1
               1
                  1
                    1
 1
   1
     1
       1
         1
           1
             1
               1
                 1
                  1
                    1
                      1
 1
     1
       1
         1
           1
             1
               1
                 1
                  1
                    1
                      1
   1
   1
     1
       1
         1
           1
             1
               1]
1 4
sneezing frequently
***************
wants somebody to cheer her up
wants somebody to cheer her up
['<s>', ' wants', ' somebody', ' to', ' cheer', ' her', ' up', '</s>']
    0 1072 4909
         7 9450
                 2 1313
                    2
Input ids [
            69
             62
               2
                       1
                         1
                      1
      1
                 1
 1
  1
    1
       1
         1
          1
            1
              1
               1
                   1
                    1
                      1
 1
  1
    1
      1
       1
         1
          1
            1
               1
                 1
                   1
                    1
              1
                      1
 1
  1
      1
       1
         1
          1
            1
               1
                 1
                    1
    1
              1
 1
  1
    1
      1
       1
         1
          1
            1
              1
               1
                 1
                   1
 1
      1
       1
         1
          1
            1
               1
                 1
                  1
  1
    1
              1
                    1
                      1
 1
  1
    1
      1
       1
         1
          1
            1
              1
               1
                 1
                  1
                    1
                      1
 1
  1
      1
       1
         1
          1
            1
              1
               1
                 1
                   1
                    1
                      1
    1
 1
  1
    1
      1
       1
         1
          1
            1
              1
               1
                 1
                   1
                    1
                      1
 1
  1]
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

```
wants somebody to cheer her up
****************
time to rinse out my conditoner
time to rinse out my conditoner
['<s>', ' time', ' to', ' rinse', ' out', ' my', ' cond', 'it', 'oner', '</s>']
Input ids [
        86
          7 39448
               66
                 127 10022
                      405 9604
                           2
                              2 7974
  2
      1
         1
           1
              1
                1
                  1
                     1
                       1
                          1
                            1
 1
      1
         1
              1
                1
                  1
                     1
                       1
                          1
                            1
    1
           1
 1
      1
         1
           1
              1
                1
                  1
                     1
                       1
                          1
                            1
 1
    1
      1
         1
           1
              1
                1
                     1
                       1
                          1
                            1
                  1
 1
    1
      1
         1
           1
              1
                1
                          1
                  1
 1
    1
      1
         1
           1
              1
                1
                  1
                     1
                       1
                          1
                            1
                            1
 1
    1
      1
         1
           1
              1
                1
                  1
                     1
                          1
                            1
 1
    1
      1
         1
           1
              1
                1
                     1
                       1
                          1
                  1
 1
    1
      1
         1
           1
              1
                1
                  1
                       1
                          1
                            1
 1
    1
      1
         1
           1
              1
                1
                  1]
end
  1 8
time to rinse out my conditoner
***************
i wasn't given an exact date, but it shouldn't be much longer. please accept our ap
please accept our apologies for the inconvenience!
     'wasn', '`', 't', 'given', 'an', 'exact', 'date',
['<s>', ' i',
                             ' but', '
         938 12905
                 576
                    41
                     6089
Input ids [
      0
       939
               90
                       1248
                           6
                                24
4395 12905
                 2540
                    3264
      90
        28
          203
            1181
                4
                       84 27620
                            13
 5 24109
      328
         2
           2
            2430
                2
                  1
                     1
                       1
                          1
                            1
 1
         1
              1
                1
                       1
                            1
    1
      1
           1
                  1
                     1
                          1
 1
    1
      1
         1
              1
                1
                     1
                            1
           1
                  1
                       1
                          1
 1
    1
      1
         1
           1
              1
                1
                  1
                     1
                       1
                          1
                            1
 1
    1
      1
         1
           1
              1
                1
                  1
                     1
                       1
                          1
                            1
 1
    1
      1
         1
           1
              1
                1
                  1
                     1
                       1
                          1
                            1
 1
      1
         1
              1
                1
                     1
                       1
                          1
                            1
    1
           1
                  1
 1
    1
      1
         1
           1
              1
                1
                  1
                     1
                       1
                          1
                            1
    1
      1
         1
           1
              1
                1
                  1]
Start tokens [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

```
19 26
please accept our apologies for the inconvenience!
*****************
hehe sure thing
hehe sure thing
['<s>', 'he', 'he',
       ' sure', ' thing', '</s>']
Input ids [
         686
     0
      37
       700
           631
              2
                2 7974
                   2
                       1
                         1
                           1
                            1
                     1
 1
   1
     1
      1
        1
          1
            1
              1
                1
                 1
                   1
                     1
                       1
                         1
 1
   1
     1
      1
        1
          1
            1
              1
                1
                 1
                   1
                     1
                       1
                         1
 1
   1
     1
      1
        1
          1
            1
              1
                 1
                   1
                     1
                       1
                         1
 1
   1
     1
      1
        1
          1
            1
              1
                1
                 1
                   1
                     1
                       1
                         1
 1
   1
     1
      1
        1
          1
            1
              1
                1
                 1
                   1
                     1
                         1
 1
   1
     1
      1
        1
          1
            1
              1
                1
                 1
                   1
                     1
                       1
                         1
 1
   1
     1
      1
        1
          1
            1
              1
                1
                 1
                   1
                       1
                         1
 1
   1
     1
      1
        1
          1
            1
              1
                1
                 1
                   1
                     1
                       1
                         1
 1
   11
1 4
hehe sure thing
****************
the cuckoo is back! this year's model starts chirping at 11 am instead of 4 am.
the cuckoo is back! this year's model starts chirping at 11 am instead of 4 am.
['<s>', ' the', ' c', 'uck', 'oo', ' is', ' back', '!', ' this', ' year', '`', 's',
Input ids [
       5
         740
          5858 3036
                16
                 124
                    328
                      42
                        76 12905
     a
1421 2012
     1855
       853 15341
           1437
                365
                  524
                    1386
                       9
                         204
              23
         7974
 524
      2
        2
                         1
    4
            2
               1
                 1
                   1
                     1
                       1
 1
    1
      1
        1
          1
            1
               1
                 1
                   1
                     1
                       1
                         1
 1
      1
        1
          1
            1
              1
                   1
                     1
                         1
    1
                 1
                       1
 1
   1
      1
        1
          1
            1
              1
                 1
                   1
                     1
                       1
                          1
 1
    1
      1
        1
          1
            1
               1
                 1
                   1
                     1
                       1
                          1
 1
    1
      1
        1
          1
            1
               1
                   1
                     1
                       1
                         1
                 1
 1
    1
      1
        1
          1
            1
               1
                 1
                   1
                     1
                       1
                         1
 1
    1
      1
        1
            1
               1
                   1
                     1
                       1
                         1
          1
                 1
        1
          1
            1
               1
                 1]
      1
```

```
1 24
     the cuckoo is back! this year's model starts chirping at 11 am instead of 4 am.
MAX LEN=128
count = y val.shape[0]
input_ids_val = np.zeros((count,MAX_LEN),dtype='int32')
attention_mask_val = np.zeros((count,MAX_LEN),dtype='int32')
start_tokens_val = np.zeros((count,MAX_LEN),dtype='int32')
end_tokens_val = np.zeros((count,MAX_LEN),dtype='int32')
from tqdm import tqdm
for i,each in tqdm(enumerate(x val.values)):
 val = tokenizer.encode_plus(each[0],each[1],add_special_tokens=True,max_length=128,retur
  input_ids_val[i] = val['input_ids']
  attention_mask_val[i] = val['attention_mask']
     5496it [00:02, 2639.78it/s]
from tqdm import tqdm
for k in tqdm(range(x_val.shape[0])):
 text1 = " "+" ".join(x_val['text'].values[k].split())
 text2 = " ".join(y_val.values[k].split())
  idx = text1.find(text2)
  chars = np.zeros((len(text1)))
  chars[idx:idx+len(text2)]=1
  if text1[idx-1]==' ':
   chars[idx-1] = 1
 enc = tokenizer.encode(text1,add_special_tokens=False)
 offsets = []; idx=0
 for t in enc:
   w = tokenizer.decode([t])
   offsets.append((idx,idx+len(w)))
   idx += len(w)
 toks = []
 for i,(a,b) in enumerate(offsets):
    sm = np.sum(chars[a:b])
    if sm>0:
      toks.append(i)
  if len(toks)>0:
    start_tokens_val[k,toks[0]+1] = 1
    end_tokens_val[k,toks[-1]+1] = 1
     100%| 5496/5496 [00:02<00:00, 2223.01it/s]
#Visualize the results
import random
for in range(35,50):
  i = random.randint(0,x_val.shape[0])
  print(x_val.iloc[i]['text'])
  print(y_val.values[i])
```

```
print([tokenizer.decode([x]) for x in tokenizer.encode(x_val.iloc[i]['text'])])
print('Input ids',input_ids_val[i])
print('attention mask',attention_mask_val[i])
print('Start tokens',start_tokens_val[i])
print('end Tokens',end_tokens_val[i])
a = np.argmax(start_tokens_val[i])
b = np.argmax(end_tokens_val[i])
print(a,b)
text1 = " "+" ".join(x_val.iloc[i]['text'].split())
enc = tokenizer.encode(text1)
st = tokenizer.decode(enc[a:b+1])
print(st)
print('*'*50)
```

```
waching gray`s antomy
waching gray`s antomy
         ' gray', '`', 's', ' ant', 'omy',
   ' W',
     'aching',
                     '</s>']
Input ids [
       885 13341 12339 12905
                29 9876 13604
                          2 7974
                              2
 1
    1
      1
        1
          1
             1
               1
                 1
                    1
                        1
                          1
 1
    1
      1
        1
          1
             1
               1
                 1
                    1
                      1
                        1
                          1
 1
    1
      1
        1
          1
             1
               1
                 1
                    1
                      1
                        1
                          1
 1
      1
               1
    1
        1
          1
             1
                 1
                    1
                      1
                        1
                          1
 1
    1
      1
        1
             1
               1
                          1
 1
    1
      1
        1
               1
                          1
          1
             1
                 1
                    1
                      1
                        1
 1
    1
      1
        1
          1
             1
               1
                 1
                    1
                      1
                        1
                          1
 1
      1
        1
             1
               1
                    1
                      1
                        1
                          1
    1
          1
                 1
        1
                        1
                          1
 1
    1
      1
          1
                 1
 1
    1
      1
        1
          1
             1
               1
                 1]
end
  1 7
waching gray`s antomy
you are right, enjoy the lasagne,
   'you', 'are', 'right', ',', 'enjoy', 'the', 'las', 'agne', ','
['<s>',
                             '</s>'
Input ids [
     0
       47
         32
           235
              6 2254
                   5 5573 16784
                          6
1313
        1
                      1
    2
      1
          1
             1
               1
                 1
                    1
                        1
                          1
 1
      1
        1
          1
             1
               1
                 1
                    1
                        1
 1
    1
      1
        1
          1
             1
               1
                    1
                      1
                        1
                          1
                 1
        1
             1
               1
 1
    1
      1
          1
                 1
 1
    1
      1
        1
          1
             1
               1
                 1
                    1
                      1
                        1
                          1
 1
      1
        1
             1
               1
                    1
                          1
    1
          1
                 1
                      1
                        1
 1
    1
      1
        1
          1
             1
               1
                 1
                    1
                      1
                        1
                          1
 1
    1
      1
        1
          1
             1
               1
                 1
                    1
                      1
                        1
                          1
 1
    1
      1
        1
          1
             1
               1
                 1
                    1
                      1
                        1
                          1
    1
      1
        1
          1
             1
               1
                 1]
5 5
enjov
  **************
with 14mph winds!!
with 14mph winds!!
['<s>', ' with', ' 14', 'mph', ' winds', '!!', '</s>']
```

```
Input ids [
        19
          501 17055 2372 12846
                      2
                              2
                                   1
            1
                 1
                    1
  1
    1
       1
          1
               1
                       1
                         1
                            1
                              1
  1
    1
       1
          1
            1
               1
                 1
                    1
                       1
                         1
                            1
                              1
  1
    1
       1
          1
               1
                 1
                       1
                         1
                              1
            1
                    1
                            1
  1
    1
       1
          1
            1
               1
                 1
                    1
                       1
                         1
                            1
                              1
  1
    1
       1
          1
            1
               1
                 1
                    1
                       1
                            1
                              1
  1
    1
       1
          1
               1
                 1
                       1
                         1
                            1
                              1
            1
                    1
  1
    1
       1
          1
            1
               1
                 1
                       1
                         1
                            1
                              1
                    1
       1
          1
               1
                 1
                       1
                         1
                            1
                              1
  1
    1
            1
                    1
          1
                       1
                            1
       1
            1
               1
                 1
                    1
                              1
  1
    1
       1
          1
            1
               1
                    1]
                 1
1 5
with 14mph winds!!
   ***************
thank you saw your comment on the chew aerobics, really tickled me
thank you saw your comment on the chew aerobics, really tickled me
['<s>', ' thank', ' you', ' ', ' saw', ' your', ' comment', ' on', ' the', ' chew',
                  110 1129
      0 3392
Input ids [
           47 1437
                794
                        15
                           5 34431 16482 2413
2857
      269 10457
           1329
                 2
                     7974
                              1
    6
              162
                    2
                         2
                            1
    1
       1
          1
               1
                 1
                    1
                       1
                              1
  1
            1
  1
    1
       1
          1
            1
               1
                 1
                    1
                       1
                         1
                            1
                              1
       1
          1
               1
                 1
                       1
                         1
                            1
  1
    1
            1
                    1
       1
          1
                 1
                         1
                            1
                              1
  1
    1
            1
               1
                    1
                       1
  1
    1
       1
          1
            1
               1
                 1
                    1
                       1
                         1
                            1
                              1
  1
       1
          1
            1
               1
                 1
                       1
                         1
                              1
    1
                    1
                            1
  1
    1
       1
          1
            1
               1
                 1
                    1
                       1
                         1
                            1
                              1
                       1
  1
    1
       1
          1
            1
               1
                 1
                    1
                         1
                            1
                              1
       1
          1
            1
               1
                 1
                    1]
1 16
thank you saw your comment on the chew aerobics, really tickled me
it`s teddy chucking time
chucking
   ' it', '`', 's', ' t', 'eddy', ' ch', 'ucking', ' time', '</s>']
      0
        24 12905
             29
                326 25561 1855 24029
Input ids [
                          86
                              2
                                2
                                 2430
  2
    1
       1
          1
            1
               1
                 1
                    1
                       1
                         1
                            1
                              1
          1
                 1
  1
    1
       1
            1
               1
                    1
                       1
                         1
                            1
                              1
            1
               1
                 1
                    1
```

```
1
    1
      1
         1
           1
             1
                1
                  1
                    1
                       1
                         1
                           1
 1
      1
         1
           1
             1
                1
                  1
                    1
                       1
                         1
                           1
    1
 1
    1
      1
         1
             1
                1
                    1
                       1
                         1
                           1
           1
                  1
 1
    1
      1
         1
           1
             1
                1
                  1
                    1
                       1
                         1
                           1
                           1
 1
    1
      1
         1
           1
             1
                1
                  1
                    1
                       1
                         1
         1
             1
                    1
                         1
                           1
      1
           1
                1
                  1
    1
      1
         1
           1
             1
                1
                  1]
6 7
chucking
   *************
at thee mall
at thee mall
['<s>', ' at', ' thee', ' mall', '</s>']
     0
       23 43419 9367
                  7974
Input ids [
                 2
                      2
                        1
                           1
                             1
                               1
         1
      1
           1
             1
                1
                    1
                       1
                           1
 1
    1
                  1
                         1
 1
      1
         1
             1
                1
                    1
                       1
                         1
                           1
    1
           1
                  1
 1
      1
         1
             1
                1
                    1
                           1
    1
           1
                  1
                       1
                         1
 1
    1
      1
         1
           1
             1
                1
                    1
                       1
                         1
                  1
                           1
 1
      1
                1
                           1
    1
         1
           1
             1
                  1
                    1
                       1
                         1
 1
    1
      1
         1
           1
             1
                1
                  1
                    1
                       1
                         1
                           1
 1
      1
         1
           1
             1
                1
                    1
                       1
                         1
                           1
    1
                  1
 1
    1
      1
         1
           1
             1
                1
                  1
                    1
                       1
                         1
                           1
 1
                       1
                           1
    1
      1
         1
           1
             1
                1
                  1
                    1
                         1
      1
         1
                1
                  1]
    1
           1
             1
1 3
at thee mall
    **************
lol - birmingham was my 1st love... but it's time to move on! ;) good morning
birmingham was my 1st love.
['<s>', 'lol', '-', 'bir', 'mingham', 'was', 'my', '1', 'st', 'love', '...',
     0 29784
         111 23629 42754
Input ids [
                 21
                   127
                     112
                       620
                          657
                            734
 24 12905
                 328 48306
                           2
      29
        86
           7
             517
               15
                      205
                        662
 2
   1313
      2
         1
           1
             1
                1
                  1
                    1
                       1
                         1
                           1
 1
      1
         1
             1
                    1
                         1
                           1
    1
           1
                1
                  1
                       1
 1
    1
      1
         1
           1
             1
                1
                  1
                    1
                       1
                         1
                           1
 1
    1
      1
         1
             1
                1
                         1
                           1
           1
                  1
                    1
                       1
      1
         1
                           1
```

```
1
       1
          1
               1
                  1
                    1
                       1
                          1
                            1
                               1
    1
            1
  1
    1
       1
          1
            1
               1
                  1
                    1
                       1
                          1
                            1
                               1
          1
               1
                  1
                    1]
            1
birmingham was my 1st love...
lmoa! i just quit one of mine, too much stress
too much stress
['<s>', ' l', 'mo', 'a', '!', ' i', ' just', ' quit', ' one', ' of', ' mine', ',', '
      0 784 4992
                   95 6602
Input ids [
           102
             328 939
                       65
                          9 4318
                               350 203
3992
      2 2430
          2
              1
                   1
                     1
                          1
    2
            1
                 1
                       1
                            1
                              1
 1
    1
      1
        1
          1
            1
               1
                 1
                   1
                     1
                       1
                            1
                              1
 1
    1
        1
          1
               1
                 1
                   1
                     1
                          1
                            1
      1
            1
                       1
                              1
 1
      1
        1
          1
            1
               1
                 1
                   1
                     1
                       1
 1
        1
               1
                 1
                   1
                     1
                       1
                          1
    1
      1
          1
            1
                            1
                              1
               1
                     1
 1
    1
      1
        1
          1
            1
                 1
                   1
                       1
                              1
 1
    1
      1
        1
          1
            1
               1
                 1
                   1
                     1
                       1
                          1
                            1
                              1
 1
    1
        1
          1
            1
               1
                 1
                   1
                     1
                       1
                          1
                            1
                              1
 1
    1]
12 14
too much stress
*************
you remind me so much of a omaha girl, that i use to date. i guess it all in the j
you remind me so much of a omaha girl, that i use to date. i guess it all in the je
['<s>', ' you', ' remind', ' me', ' so', ' much', ' of', ' a', ' om', 'aha', ' girl'
Input ids [
      0
        47
          8736
             162
                98
                   203
                      9
                        10 16780 11695
                               1816
          7
 14
   939
           1248
               4
                1437
                   939
                      4443
                         24
                            70
      304
                              11
                   459
  5 10844
       6
         36 11762
              6320
                 449
                      1344
                         16
                            99
                               79
       2
          2
           7974
5328
    322
               2
                  1
                    1
                       1
                          1
                            1
                               1
  1
       1
          1
            1
               1
                  1
                    1
                       1
                            1
                               1
    1
                          1
  1
    1
       1
          1
            1
               1
                  1
                    1
                       1
                          1
                            1
                               1
  1
    1
       1
          1
            1
               1
                  1
                    1
                       1
                          1
                            1
                               1
  1
    1
       1
          1
            1
               1
                  1
                    1
                       1
                          1
                            1
                               1
  1
    1
       1
          1
               1
                  1
                       1
                          1
                               1
            1
                    1
                            1
  1
    1
       1
          1
            1
               1
                  1
                    1
                       1
                          1
                            1
                               1
                  1
                    1]
    1
       1
          1
            1
               1
```

```
you remind me so much of a omaha girl, that i use to date. i guess it all in the je
*****************
the palm pre doesn`t record video
the palm pre doesn't record video
['<s>', ' the', ' palm', ' pre', ' doesn', '`', 't', ' record', ' video', '</s>']
Input ids [
        5 14262 1198
              630 12905
                    90
                      638
                        569
                           2
                              2 7974
  2
    1
      1
         1
           1
              1
                1
                   1
                     1
                       1
                          1
                            1
  1
      1
         1
           1
              1
                1
                   1
                     1
                       1
                          1
                            1
    1
  1
      1
         1
           1
              1
                1
                   1
                     1
                          1
                            1
  1
    1
      1
         1
              1
                1
                     1
                            1
           1
                   1
                       1
                          1
  1
    1
      1
         1
           1
              1
                1
                   1
      1
                            1
  1
    1
         1
           1
              1
                1
                   1
                     1
                       1
                          1
         1
                     1
                            1
  1
    1
      1
           1
              1
                1
                   1
                            1
  1
    1
      1
         1
           1
              1
                1
                     1
                       1
                          1
                   1
  1
    1
      1
         1
           1
              1
                1
                   1
                       1
                          1
                            1
  1
    1
      1
         1
           1
              1
                1
                   1]
end
1 8
the palm pre doesn't record video
****************
daarling haha yah it`s walking distance from my house mission tiki drive-in haha e
_daarling haha yah it`s walking distance from my house mission tiki drive-in haha e
    _', 'da', 'arling', ' haha', ' y', 'ah', ' it', '`', 's', ' walking', ' di
Input ids [
      0 18134 6106 34529 46116 1423 895
                      24 12905
                           29
                             3051 4472
                 1305
                       179 46116
 31
   127
      790 1437
          2511
             326 8907
                     12
                           358
        101
                    7974
1035
    52
      213
          358
             983
                2
                   2
                       2
                          1
                            1
      1
         1
              1
                1
                   1
                       1
                          1
                            1
  1
    1
           1
                     1
  1
    1
      1
         1
           1
              1
                1
                     1
                            1
                   1
                       1
                          1
  1
    1
      1
         1
           1
              1
                1
                   1
                     1
                       1
                          1
                            1
  1
    1
      1
         1
           1
              1
                1
                   1
                     1
                       1
                          1
                            1
  1
    1
      1
         1
           1
              1
                1
                   1
                     1
                       1
                          1
                            1
  1
      1
         1
              1
                1
                     1
                       1
                            1
    1
           1
                   1
                          1
  1
    1
      1
         1
           1
              1
                1
                   1
                     1
                       1
                          1
                            1
      1
         1
           1
              1
                   1]
  1
    1
                1
```

```
_daarling haha yah it`s walking distance from my house mission tiki drive-in haha e
*****************
i`m having a fantastically miserable time so far...
fantastically miserable time
['<s>', ' i', '`', 'm', ' having', ' a', ' fant', 'astically', ' miserable', ' time'
Input ids [
      0
       939 12905
            119
              519
                 10 30934 41936 20161
                               444
      2 7974
           2
 734
    2
              1
                1
                  1
                     1
                       1
                          1
                            1
 1
      1
         1
              1
                1
                  1
                     1
                       1
                          1
                            1
    1
           1
 1
    1
      1
         1
           1
              1
                1
                  1
                     1
                       1
                          1
                            1
 1
    1
      1
         1
           1
              1
                1
                  1
                     1
                       1
                          1
                            1
 1
    1
      1
         1
           1
              1
                1
                   1
                     1
                       1
                          1
 1
    1
      1
         1
           1
              1
                1
                  1
                     1
                       1
                          1
                            1
 1
         1
              1
                            1
    1
      1
           1
                1
                  1
                     1
                          1
                     1
                            1
 1
    1
      1
         1
           1
              1
                1
                       1
                          1
                  1
 1
    1
      1
         1
           1
              1
                1
                  1
                     1
                       1
                          1
                            1
 1
    1
      1
         1
           1
              1
                1
                  1]
attention mask [1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
end
  6 9
fantastically miserable time
**************
i don't use it, but i heard that powershell is nice
['<s>', ' i', ' don', '`', 't', ' use', ' it', ',', ' but', ' i', ' heard', ' that',
       939
          218 12905
Input ids [
      0
               90
                 304
                    24
                      6
                        53
                            1317
       2579
4361 20030
              2
                  2
      16
           2
               1313
                     1
                       1
                          1
                            1
 1
    1
      1
         1
           1
              1
                1
                  1
                     1
                       1
                          1
                            1
 1
      1
         1
              1
                1
                     1
                          1
                            1
    1
           1
                  1
                       1
 1
    1
      1
         1
           1
              1
                1
                  1
                     1
                       1
                          1
                            1
 1
    1
      1
         1
           1
              1
                1
                  1
                     1
                       1
                          1
                            1
 1
         1
           1
                1
                          1
                            1
    1
      1
              1
                  1
                     1
                       1
 1
    1
      1
         1
           1
              1
                1
                   1
                     1
                       1
                          1
                            1
 1
    1
      1
         1
           1
              1
                1
                     1
                       1
                          1
                            1
                  1
 1
    1
      1
         1
           1
              1
                1
                  1
                     1
                       1
                          1
                            1
    1
      1
         1
           1
              1
                1
                  1]
```

```
end
   15 15
   nice
  **************
  thanks for the props
  thanks for the props
  ['<s>', ' thanks', ' for', ' the', ' props',
                         '</s>']
  Input ids [
          0 2446
                   5 26504
                          2
                13
                             2 1313
                                   2
                                      1
                                         1
                                            1
     1
        1
           1
              1
                 1
                    1
                       1
                           1
                              1
                                 1
                                    1
                                       1
     1
        1
           1
              1
                    1
                       1
                                 1
                                       1
                 1
                           1
                              1
                                    1
     1
        1
           1
              1
                 1
                    1
                       1
                           1
                              1
                                 1
                                    1
                                       1
     1
        1
           1
              1
                 1
                    1
                       1
                           1
                              1
                                 1
                                    1
                                       1
     1
        1
           1
              1
                 1
                    1
                       1
                           1
                              1
                                 1
                                    1
                                       1
     1
        1
           1
              1
                 1
                    1
                       1
                           1
                              1
                                 1
                                    1
                                       1
     1
        1
           1
              1
                 1
                    1
                       1
                           1
                              1
                                    1
                                       1
                                       1
     1
        1
           1
              1
                 1
                    1
                       1
                              1
                                 1
                                    1
                           1
     1
        1
           1
              1
                 1
                    1
                       1
                           1
                              1
                                 1
                                    1
                                       1
     1
           1
              1
                           1]
        1
                 1
                    1
                       1
  end
     1 4
   thanks for the props
  ***************
  i hope i can find a store to take us
  i hope i can find a store to take us
  ['<s>', ' i', ' hope', ' i', ' can', ' find', ' a', ' store', ' to', ' take', ' us',
  Input ids [
         0 939 1034 939
                     465
                        10 1400
                              7 185 201
                   64
from transformers import TFRobertaForQuestionAnswering
roberta = TFRobertaForQuestionAnswering.from pretrained('roberta-base')
  Some layers from the model checkpoint at roberta-base were not used when initializin
  - This IS expected if you are initializing TFRobertaForQuestionAnswering from the ch
  - This IS NOT expected if you are initializing TFRobertaForQuestionAnswering from th
  Some layers of TFRobertaForQuestionAnswering were not initialized from the model che
  You should probably TRAIN this model on a down-stream task to be able to use it for
```

+f aveand dime/dean1 avic

```
arop1 = tT.expana_aims(arop1,axis=-1)
```

layer1 = tf.keras.layers.Conv1D(1,1)(drop1)

layer1= Flatten()(layer1)

#dense1 = Dense(units=MAX_LEN,activation='relu',name='dense1',kernel_regularizer = tf.kera

softmax1 = Activation('softmax')(layer1)

#dense2 = Dense(units=MAX_LEN,activation='relu',name='dense2',kernel_regularizer = tf.kera

drop2 = Dropout(0.1)(end_scores)

drop2 = tf.expand_dims(drop2,axis=-1)

layer2 = tf.keras.layers.Conv1D(1,1)(drop2)

layer2 = Flatten()(layer2)

softmax2 = Activation('softmax')(layer2)

model = Model(inputs=[input1,input2],outputs=[softmax1,softmax2])

model.summary()

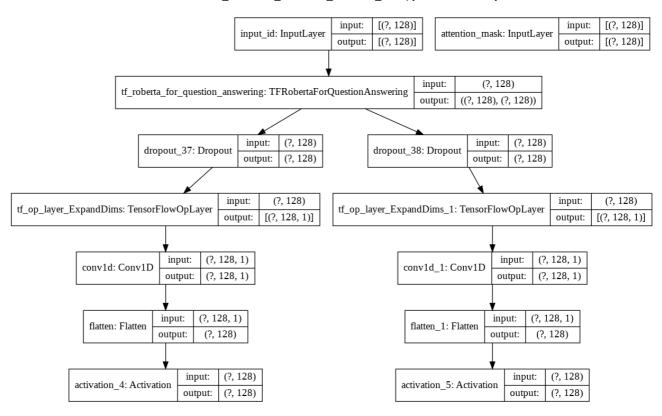
Model: "functional_1"

Layer (type)	Output Shape	Param #	Connected to
input_id (InputLayer)	[(None, 128)]	0	
attention_mask (InputLayer)	[(None, 128)]	0	
tf_roberta_for_question_answeri	((None, 128), (None,	124647170	<pre>input_id[0][0] attention_mask[0][0</pre>
dropout_37 (Dropout)	(None, 128)	0	tf_roberta_for_ques
dropout_38 (Dropout)	(None, 128)	0	tf_roberta_for_ques
tf_op_layer_ExpandDims (TensorF	[(None, 128, 1)]	0	dropout_37[0][0]
tf_op_layer_ExpandDims_1 (Tenso	[(None, 128, 1)]	0	dropout_38[0][0]
conv1d (Conv1D)	(None, 128, 1)	2	tf_op_layer_ExpandD
conv1d_1 (Conv1D)	(None, 128, 1)	2	tf_op_layer_ExpandD
flatten (Flatten)	(None, 128)	0	conv1d[0][0]
flatten_1 (Flatten)	(None, 128)	0	conv1d_1[0][0]
activation_4 (Activation)	(None, 128)	0	flatten[0][0]
activation_5 (Activation)	(None, 128)	0	flatten_1[0][0]

Total params: 124,647,174
Trainable params: 124,647,174

Non-trainable params: 0

```
import tensorflow as tf
tf.keras.utils.plot_model(model, 'Model.png',show_shapes=True)
```



```
! rm -r '/content/checkpt1'
! rm -r '/content/tensorboard_logs1'

input_data = (input_ids,attention_mask)
output_data = (start_tokens,end_tokens)

val = (input_ids_val,attention_mask_val)
output_val = (start_tokens_val,end_tokens_val)
val_data = (val,output_val)

%load_ext tensorboard
import datetime
import os
//colab research google com/drive/1eOH6gaArm1i3TwDE2dulbr34XIMROfCi
```

```
#reduce lr = tf.keras.callbacks.ReduceLROnPlateau(monitor='val loss', factor=0.1,patience=
log\_dir=\ os.path.join("tensorboard\_logs1"\ ,\ datetime.datetime.now().strftime("%Y%m%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%MM%d-%H%M%d-%H%MMM%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-%H%M%d-
tensorboard callback = tf.keras.callbacks.TensorBoard(log dir=log dir,histogram freq=1, wr
! mkdir 'checkpt1'
file_path = os.path.join('checkpt1/model_roberta.hdf5')
checkpt_save = tf.keras.callbacks.ModelCheckpoint(filepath=file_path,save_weights_only=Tru
callbacks=[tensorboard_callback,checkpt_save]
opt = tf.keras.optimizers.Adam(learning rate=3e-5, epsilon=1e-08, clipnorm=1.0)
model.compile(optimizer=opt,loss='categorical_crossentropy')
train_dataset = tf.data.Dataset.from_tensor_slices((input_data, output_data)).shuffle(buff
val_dataset = tf.data.Dataset.from_tensor_slices(val_data).batch(32)
model.fit(train dataset,epochs=6,validation data=val dataset,callbacks=callbacks)
       WARNING:tensorflow:Model failed to serialize as JSON. Ignoring...
       Epoch 1/6
       WARNING:tensorflow:Gradients do not exist for variables ['tf_roberta_for_question_an
          1/687 [.....] - ETA: 0s - loss: 9.4676 - activation_4_los
       Instructions for updating:
       use `tf.profiler.experimental.stop` instead.
       687/687 [============= ] - ETA: 0s - loss: 3.0323 - activation_4_los
       Epoch 00001: val_loss improved from inf to 1.68423, saving model to checkpt1/model_r
       687/687 [============] - 675s 982ms/step - loss: 3.0323 - activati
       Epoch 2/6
       Epoch 00002: val_loss improved from 1.68423 to 1.64433, saving model to checkpt1/mod
       Epoch 3/6
       Epoch 00003: val_loss improved from 1.64433 to 1.61733, saving model to checkpt1/mod
       687/687 [=============== ] - 672s 978ms/step - loss: 2.0634 - activati
       Epoch 4/6
       Epoch 00004: val_loss did not improve from 1.61733
       687/687 [========================== ] - 667s 971ms/step - loss: 1.9122 - activati
       Epoch 5/6
       Epoch 00005: val loss did not improve from 1.61733
       687/687 [========================== ] - 667s 971ms/step - loss: 1.7760 - activati
       Epoch 6/6
       Epoch 00006: val_loss did not improve from 1.61733
       687/687 [============ ] - 668s 972ms/step - loss: 1.6315 - activati
       <tensorflow.python.keras.callbacks.History at 0x7f8e8e7172e8>
```

For training data

```
start , end = model.predict(input_data)
start.shape,end.shape
     ((21984, 128), (21984, 128))
pred_values=[]
from tqdm import tqdm
for i in tqdm(range(start.shape[0])):
 a = np.argmax(start[i])
 b = np.argmax(end[i])
 text1 = " "+" ".join(x_train['text'].values[i].split())
 enc = tokenizer.encode(text1)
 val = tokenizer.decode(enc[a:b+1])
 pred_values.append(val)
     100%| 21984/21984 [00:05<00:00, 4376.17it/s]
x_train['pred_text'] = pred_values
x_train['selected_text'] = y_train.values
def jaccard(str1, str2):
    a = set(str1.lower().split())
   b = set(str2.lower().split())
   if (len(a)==0) & (len(b)==0): return 0.5
   c = a.intersection(b)
    return float(len(c)) / (len(a) + len(b) - len(c))
scores=[]
for i in tqdm(range(x_train.shape[0])):
  scores.append(jaccard(x_train['pred_text'].values[i],x_train['selected_text'].values[i])
x_train['jaccard']=scores
     100%| 21984/21984 [00:00<00:00, 106555.12it/s]
x_train.head(4)
```

	text	sentiment	pred_text	selected_text	jaccard
21209	oh, he`s hilarious. i`m just commenting on th	positive	hilarious.	hilarious.	1.000000
18265	thanks for trying i was hoping bud trillin,	positive	thanks	thanks for trying i was hoping bud trillin,	0.045455

```
print('Mean Jaccard score for training data:',x_train['jaccard'].values.mean())
```

Mean Jaccard score for training data: 0.7618829188841084

For Validation data

```
val = (input_ids_val,attention_mask_val)
start_val , end_val = model.predict(val)
start_val.shape,end_val.shape
     ((5496, 128), (5496, 128))
pred_values_val=[]
from tqdm import tqdm
for i in tqdm(range(start_val.shape[0])):
  a = np.argmax(start_val[i])
 b = np.argmax(end_val[i])
 text1 = " "+" ".join(x_val['text'].values[i].split())
 enc = tokenizer.encode(text1)
 val = tokenizer.decode(enc[a:b+1])
 pred_values_val.append(val)
     100% | 5496/5496 [00:01<00:00, 4360.44it/s]
scores_val=[]
x_val['pred_text'] = pred_values_val
x_val['selected_text'] = y_val.values
scores=[]
for i in tqdm(range(x val.shape[0])):
  scores_val.append(jaccard(x_val['pred_text'].values[i],x_val['selected_text'].values[i])
x_val['jaccard']=scores_val
     100%| 5496/5496 [00:00<00:00, 104193.59it/s]
x_val.sample(20)
```

	text	sentiment	pred_text	selected_text	jaccard
8280	they raised the prices at work which means peo	negative	less.	raised the prices	0.000000
10043	it shall be mine!! it shall be mine!!! *ra	neutral	it shall be mine!! it shall be mine!!! *ra	it shall be mine!! it shall be mine!!! *rai	1.000000
17078	what a smashing show tonight! love the killer	neutral	what a smashing show tonight! love the killer	what a smashing show tonight! love the killers	1.000000
283	happy mothers day people. i love my mom a lot	positive	happy	happy mothers day people. i love my mom a lot	0.090909
26748	i want to panic	neutral	i want to panic	i want to panic	1.000000
12015	my bet on mumbai indians today anyone to ta	neutral	my bet on mumbai indians today anyone to t	my bet on mumbai indians today anyone to ta	1.000000
17584	big header folio (http://bit.ly/mymea has bee	negative	i`m really impatient	i`m really impatient lol	0.750000
	hought awesome			amesome	

print('Mean Jaccard score for val data:',x_val['jaccard'].values.mean())

Mean Jaccard score for val data: 0.7009884376335759

```
print(x_val[x_val.sentiment=='positive']['jaccard'].values.mean())
print(x_val[x_val.sentiment=='negative']['jaccard'].values.mean())
print(x_val[x_val.sentiment=='neutral']['jaccard'].values.mean())
```

- 0.5194540529961196
- 0.5221797725160238
- 0.9637421329970904

Error Analysis

x_val.head()

	text	sentiment	pred_text	selected_text	jaccard
1589	enjoy! family trumps everything	positive	enjoy!	enjoy! family trumps everything	0.250000
10414	of them kinda turns me off of it all. and	negative	turns me off of it all.	kinda turns me off	0.428571
6562	clive it`s my birthday pat me http://apps.fac	neutral	clive it`s my birthday pat me	clive it`s my birthday pat me	1.000000

x_val.head()

	text	sentiment	<pre>pred_text</pre>	selected_text	jaccard	text_len	selte
1589	enjoy! family trumps everything	positive	enjoy!	enjoy! family trumps everything	0.250000	4	
10414	of them kinda turns me off of it all. and	negative	turns me off of it all.	kinda turns me off	0.428571	23	
	clive it's mv		clive it's				

print('Mean & Median values for difference in length of text and selected_text for Neutral print('Mean & Median values for difference in length of text and selected_text for Positiv print('Mean & Median values for difference in length of text and selected_text for Negativ

Mean & Median values for difference in length of text and selected_text for Neutral: Mean & Median values for difference in length of text and selected_text for Positive Mean & Median values for difference in length of text and selected_text for Negative

print('Mean Jaccard score when difference is less than 5:',x_val[x_val['diff'] <= 5]['jacc
print('Mean Jaccard score when difference is between 5 and 10:',x_val[(x_val['diff'] > 5)
print('Mean Jaccard score when difference is greater than 10:',x_val[x_val['diff'] > 10]['

```
Mean Jaccard score when difference is less than 5: 0.8216090075732144

Mean Jaccard score when difference is between 5 and 10: 0.5450486615053474

Mean Jaccard score when difference is greater than 10: 0.5059911270103421
```

Conclusion

- If you calculate the difference between the number of words in the text and selected_text
 columns for Neutral data, the mean and median values suggests that there is almost no
 difference in the no.of words in text and selected_text and the model performs well in this
 scenario
- From the above, the difference between number of words in text and selected_text
 columns for the Positive and Negative data are around 9.3 and 9.5, which suggests that
 as the difference between text and selected_text increases, the model doesn't seem to
 perform better

- The model performs well when the **difference** between the no.of words in text and selected_text columns is **small (less than 5)**, the jaccard score is around **0.81**
- When the difference value increases (greater than 5), the model performance seems to be dropping as we get a jaccard score of around **0.52**