PreProcessingPipeline

October 21, 2023

[ ]:
```
#colab
#!pip install pyspellchecker
```

[ ]:
```
import pandas as pd
import numpy as np
import string

from spellchecker import SpellChecker
import nltk
from nltk.tokenize.treebank import TreebankWordDetokenizer
from nltk.tokenize import word_tokenize

import tensorflow as tf
from tensorflow import keras
```

```
2023-10-21 09:03:12.069932: I tensorflow/tsl/cuda/cudart_stub.cc:28] Could not
find cuda drivers on your machine, GPU will not be used.
2023-10-21 09:03:12.156033: E
tensorflow/compiler/xla/stream_executor/cuda/cuda_dnn.cc:9342] Unable to
register cuDNN factory: Attempting to register factory for plugin cuDNN when one
has already been registered
2023-10-21 09:03:12.156090: E
tensorflow/compiler/xla/stream_executor/cuda/cuda_fft.cc:609] Unable to register
cuFFT factory: Attempting to register factory for plugin cuFFT when one has
already been registered
2023-10-21 09:03:12.156150: E
tensorflow/compiler/xla/stream_executor/cuda/cuda_blas.cc:1518] Unable to
register cuBLAS factory: Attempting to register factory for plugin cuBLAS when
one has already been registered
2023-10-21 09:03:12.171902: I tensorflow/tsl/cuda/cudart_stub.cc:28] Could not
find cuda drivers on your machine, GPU will not be used.
2023-10-21 09:03:12.172378: I tensorflow/core/platform/cpu_feature_guard.cc:182]
This TensorFlow binary is optimized to use available CPU instructions in
performance-critical operations.
To enable the following instructions: AVX2 FMA, in other operations, rebuild
TensorFlow with the appropriate compiler flags.
2023-10-21 09:03:14.100288: W
tensorflow/compiler/tf2tensorrt/utils/py_utils.cc:38] TF-TRT Warning: Could not
```

```
find TensorRT
WARNING:root:Limited tf.compat.v2.summary API due to missing TensorBoard
installation.
WARNING:root:Limited tf.compat.v2.summary API due to missing TensorBoard
installation.
WARNING:root:Limited tf.compat.v2.summary API due to missing TensorBoard
installation.
WARNING:root:Limited tf.summary API due to missing TensorBoard installation.
WARNING:root:Limited tf.compat.v2.summary API due to missing TensorBoard
installation.
WARNING:root:Limited tf.compat.v2.summary API due to missing TensorBoard
installation.
WARNING:root:Limited tf.compat.v2.summary API due to missing TensorBoard
installation.
```

0.1 Dataset

[ ]:
```
#colab
#from google.colab import files
#upload = files.upload()
```

[ ]:
```
#colab
#import io
#df = pd.read_csv(io.BytesIO(upload['data.csv']))
#df
```

[ ]:
```
df = pd.read_csv('./data.csv')
df
```

[ ]:

	Sentence	Sentiment
0	The GeoSolutions technology will leverage Bene…	positive
1	$ESI on lows, down $1.50 to $2.50 BK a real po…	negative
2	For the last quarter of 2010 , Componenta 's n…	positive
3	According to the Finnish-Russian Chamber of Co…	neutral
4	The Swedish buyout firm has sold its remaining…	neutral
…	…	…
5837	RISING costs have forced packaging producer Hu…	negative
5838	Nordic Walking was first used as a summer trai…	neutral
5839	According shipping company Viking Line , the E…	neutral
5840	In the building and home improvement trade , s…	neutral
5841	HELSINKI AFX - KCI Konecranes said it has won …	positive

[5842 rows x 2 columns]

0.2 Descricao

[ ]:
```
print('Shape ' + str(df.shape))
```

```
Shape (5842, 2)
```

[ ]:
```
print('is there Null?')
print(df.isnull().sum())
```

```
is there Null?
Sentence     0
Sentiment    0
dtype: int64
```

[ ]:
```
print('Distribuicao das classes')
classe = df['Sentiment'].value_counts()
classe = pd.DataFrame(classe, columns=['Sentiment'])
classe['% total'] = (classe['Sentiment'] / df.shape[0]).round(2)
classe
```

```
Distribuicao das classes
```

[ ]:
```
Empty DataFrame
Columns: [Sentiment, % total]
Index: []
```

[ ]:
```
df.groupby('Sentiment').describe()
```

[ ]:
```
          Sentence                                                                \
             count unique                                                    top
Sentiment
negative       860    860  $ESI on lows, down $1.50 to $2.50 BK a real po…
neutral       3130   3124  SSH Communications Security Corporation is hea…
positive      1852   1852  The GeoSolutions technology will leverage Bene…


          freq
Sentiment
negative     1
neutral      2
positive     1
```

0.3 PRE-PROCESSAMENTO

[ ]:
```
corpus = df
```

[ ]:
```
l_unique = sorted(corpus['Sentiment'].unique())
label_map = {sentiment: i for i, sentiment in enumerate(l_unique)}
```

```
label_map
```

[ ]:
```
{'negative': 0, 'neutral': 1, 'positive': 2}
```

[ ]:
```
corpus['Class'] = corpus['Sentiment'].map(label_map)
corpus
```

[ ]:
```
                                              Sentence Sentiment  Class
0     The GeoSolutions technology will leverage Bene…  positive      2
1     $ESI on lows, down $1.50 to $2.50 BK a real po…  negative      0
2     For the last quarter of 2010 , Componenta 's n…  positive      2
3     According to the Finnish-Russian Chamber of Co…   neutral      1
4     The Swedish buyout firm has sold its remaining…   neutral      1
…                                                   …       …    …
5837  RISING costs have forced packaging producer Hu…  negative      0
5838  Nordic Walking was first used as a summer trai…   neutral      1
5839  According shipping company Viking Line , the E…   neutral      1
5840  In the building and home improvement trade , s…   neutral      1
5841  HELSINKI AFX - KCI Konecranes said it has won …  positive      2

[5842 rows x 3 columns]
```

[ ]:
```
from nltk.stem import WordNetLemmatizer
from nltk.corpus import stopwords
```

[ ]:
```
#https://www.kaggle.com/code/jovanchua/financial-statement-analysis


# tokenizer = TreebankWordTokenizer()
# lemmatizer = WordNetLemmatizer()
# stop_words = set(stopwords.words('english'))
# tool = language_tool_python.LanguageTool('en-US')
#def preprocess_text(text, tokenizer, lemmatizer, stop_words, spellchecker):

textos = corpus['Sentence']
textos = textos.str.lower()
textos = textos.str.translate(str.maketrans('', '', string.punctuation))
textos = textos.str.replace('[\d+]', '') #remove numeros
#textos = textos.str.replace(None, '')
```

[ ]:
```
corpus['SentenceAdj'] = textos
corpus
```

[ ]:
```
                                              Sentence Sentiment  Class  \
0     The GeoSolutions technology will leverage Bene…  positive      2
1     $ESI on lows, down $1.50 to $2.50 BK a real po…  negative      0
2     For the last quarter of 2010 , Componenta 's n…  positive      2
```

```
3     According to the Finnish-Russian Chamber of Co…   neutral      1
4     The Swedish buyout firm has sold its remaining…   neutral      1
…                                                  …       …    …
5837  RISING costs have forced packaging producer Hu…  negative      0
5838  Nordic Walking was first used as a summer trai…   neutral      1
5839  According shipping company Viking Line , the E…   neutral      1
5840  In the building and home improvement trade , s…   neutral      1
5841  HELSINKI AFX - KCI Konecranes said it has won …  positive      2

                                              SentenceAdj
0     the geosolutions technology will leverage bene…
1     esi on lows down 150 to 250 bk a real possibility
2     for the last quarter of 2010  componenta s net…
3     according to the finnishrussian chamber of com…
4     the swedish buyout firm has sold its remaining…
…                                                  …
5837  rising costs have forced packaging producer hu…
5838  nordic walking was first used as a summer trai…
5839  according shipping company viking line  the eu…
5840  in the building and home improvement trade  sa…
5841  helsinki afx  kci konecranes said it has won a…

[5842 rows x 4 columns]
```

[ ]:
```
# textos = corpus[['SentenceAdj', 'Sentiment']].to_numpy()
# textos
```

[ ]:
```
#colab
#import nltk
#nltk.download('punkt')
```

0.4 MODEL

Source

[ ]:
```
import tensorflow as tf
```

[ ]:
```
from tensorflow import keras
from tensorflow.keras import Sequential
```

[ ]:
```

```

[ ]:
```
from tensorflow.keras import Sequential
```

[ ]:
```
from tensorflow.keras import layers
from tensorflow.keras.layers import TextVectorization
from sklearn.model_selection import train_test_split
```

```
import os
os.environ['TF_CPP_MIN_LOG_LEVEL'] = '2'
```

[ ]:

[ ]:
```
x_train, x_test, y_train, y_test = train_test_split(corpus['SentenceAdj'],␣
 ↪corpus['Class'], test_size=0.2, random_state=42)
print('df split done')
x_train
```

```
df split done
```

[ ]:
```
1647    the floor area of the yliopistonrinne project …
1669    no compensation for its news  opinions or dist…
3159        rt acinvestorblog aapl still on track for 500
4577    this includes a eur 395 mn change in the fair …
4221                                    gte long at 744
                              …
3772    bullya pollux654321 my 50 kors 80 calls are go…
5191    according to seppñnen  the new technology umts…
5226      crus upgraded to a buy by alpha street research
5390    favourable currency rates also contributed to …
860     tesco breaks its downward slide by cutting sal…
Name: SentenceAdj, Length: 4673, dtype: object
```

[ ]:
```
corpus
```

[ ]:
```
                                              Sentence Sentiment  Class  \
0     The GeoSolutions technology will leverage Bene…  positive      2
1     $ESI on lows, down $1.50 to $2.50 BK a real po…  negative      0
2     For the last quarter of 2010 , Componenta 's n…  positive      2
3     According to the Finnish-Russian Chamber of Co…   neutral      1
4     The Swedish buyout firm has sold its remaining…   neutral      1
…                                                   …       …    …
5837  RISING costs have forced packaging producer Hu…  negative      0
5838  Nordic Walking was first used as a summer trai…   neutral      1
5839  According shipping company Viking Line , the E…   neutral      1
5840  In the building and home improvement trade , s…   neutral      1
5841  HELSINKI AFX - KCI Konecranes said it has won …  positive      2

                                          SentenceAdj
0     the geosolutions technology will leverage bene…
1     esi on lows down 150 to 250 bk a real possibility
2     for the last quarter of 2010  componenta s net…
3     according to the finnishrussian chamber of com…
4     the swedish buyout firm has sold its remaining…
```

```
…                                                    …
5837  rising costs have forced packaging producer hu…
5838  nordic walking was first used as a summer trai…
5839  according shipping company viking line  the eu…
5840  in the building and home improvement trade  sa…
5841  helsinki afx  kci konecranes said it has won a…

[5842 rows x 4 columns]
```

[ ]:
```
import tensorflow_hub as hub
import tensorflow_text as text
```

[ ]:
```
bert_preprocess = hub.KerasLayer("https://tfhub.dev/tensorflow/
 ↪bert_en_uncased_preprocess/3")
bert_encoder = hub.KerasLayer("https://tfhub.dev/tensorflow/
 ↪bert_en_uncased_L-12_H-768_A-12/4")
```

[ ]:
```
def get_sentence_embeding(sentences):
    preprocessed_text = bert_preprocess(sentences)
    return bert_encoder(preprocessed_text)['pooled_output']
```

[ ]:
```
x_train[0]
```

[ ]:
```
'the geosolutions technology will leverage benefon s gps solutions by providing
location based search technology  a communities platform  location relevant
multimedia content and a new and powerful commercial model '
```

[ ]:
```
#emb = get_sentence_embeding(corpus['SentenceAdj'])
emb = get_sentence_embeding(['the geosolutions technology will leverage benefon␣
 ↪s gps solutions by providing location based search technology  a communities␣
 ↪platform  location relevant multimedia content and a new and powerful␣
 ↪commercial model '])
```

[ ]:
```
emb
```

[ ]:
```
<tf.Tensor: shape=(1, 768), dtype=float32, numpy=
array([[-0.7296258 , -0.19282198, -0.8115986 ,  0.64450693,  0.51574135,
        -0.2046549 ,  0.2883235 ,  0.26509354, -0.58445907, -0.999921  ,
        -0.11365996,  0.71174175,  0.89551145,  0.27836877,  0.5998207 ,
        -0.27222365,  0.10308256, -0.4413118 ,  0.3070638 ,  0.31077045,
         0.54338557,  0.9999777 ,  0.10377765,  0.3816154 ,  0.3377827 ,
         0.892626  , -0.5890046 ,  0.7371317 ,  0.8520142 ,  0.69433326,
        -0.38364512,  0.23554745, -0.949108  , -0.10279756, -0.91568464,
        -0.949397  ,  0.32838124, -0.5974247 ,  0.13797197, -0.02844633,
        -0.7040172 ,  0.34275144,  0.9997806 , -0.3200955 ,  0.52280843,
        -0.12689726, -0.99998873,  0.1061333 , -0.71831274,  0.80368805,
         0.5147016 ,  0.74162805,  0.16598816,  0.43803552,  0.36884943,
```

```
-0.27803102, -0.27399957,  0.07402167, -0.1500818 , -0.4658085 ,
-0.5079623 ,  0.37536815, -0.7514504 , -0.760566  ,  0.560081  ,
 0.6630013 , -0.11066921, -0.2606113 , -0.10052219,  0.02116172,
 0.61772937,  0.12292605, -0.16857675, -0.63106817,  0.26185593,
 0.19723043, -0.67249715,  1.        , -0.41497567, -0.8718446 ,
 0.6413123 ,  0.566339  ,  0.66346884, -0.10714938,  0.5085824 ,
-1.        ,  0.52109855, -0.10296167, -0.9476755 ,  0.16437736,
 0.4692983 , -0.0692717 ,  0.21593504,  0.6842841 , -0.10345069,
-0.48427102, -0.20046362, -0.72981733, -0.36329597, -0.47484833,
 0.15944527, -0.24376373, -0.12244146, -0.28327957,  0.28013474,
-0.30104455, -0.19848207,  0.5698105 , -0.27516097,  0.44734067,
 0.60752845, -0.31876394,  0.1347419 , -0.8255777 ,  0.5078688 ,
-0.26069397, -0.9626543 , -0.6659797 , -0.941619  ,  0.41116628,
 0.06337499, -0.07367007,  0.7230267 , -0.20893885,  0.30760613,
-0.11414724, -0.76237774, -1.        , -0.5763317 , -0.24203792,
 0.14464214, -0.17900307, -0.8882415 , -0.8701448 ,  0.5532696 ,
 0.8942184 ,  0.17753063,  0.9988072 , -0.35193795,  0.7388289 ,
-0.06322075, -0.617453  ,  0.27561584, -0.28824756,  0.8095847 ,
-0.05311031, -0.30168435,  0.19207887, -0.526752  ,  0.04159917,
-0.60060775, -0.13232012, -0.49626094, -0.7174971 , -0.2789953 ,
 0.81038845, -0.41075778, -0.7750682 ,  0.1087581 , -0.11696021,
-0.38601175,  0.76969916,  0.59321856,  0.18347795, -0.09245849,
 0.36573073, -0.13947481,  0.39797452, -0.6602406 , -0.30369526,
 0.36586362, -0.2748161 , -0.75066835, -0.8631648 , -0.23073286,
 0.39451867,  0.9292377 ,  0.5712688 ,  0.28879526,  0.42043722,
-0.37235108,  0.3737087 , -0.9206899 ,  0.9159602 , -0.06457204,
 0.1897253 , -0.01388466,  0.56860876, -0.6577584 , -0.14022298,
 0.5771474 , -0.60885006, -0.5599458 , -0.01669196, -0.46747524,
-0.42274433, -0.5866258 ,  0.5196313 , -0.19773683, -0.3136036 ,
-0.08157942,  0.7581206 ,  0.8182883 ,  0.49354553,  0.10702389,
 0.62213224, -0.60029876, -0.12968026,  0.15888646,  0.08283813,
 0.11469878,  0.9467451 , -0.5918506 , -0.01717309, -0.7893503 ,
-0.9091546 , -0.02827637, -0.7060055 , -0.03470754, -0.7278255 ,
 0.58337563, -0.5064701 ,  0.04995491,  0.30857548, -0.6028132 ,
-0.38763854,  0.0547794 , -0.44984725,  0.39480793, -0.21663503,
 0.82804084,  0.8739874 , -0.48017228, -0.24080028,  0.8954913 ,
-0.82131   , -0.65444475,  0.08819579, -0.3328971 ,  0.65252286,
-0.50323886,  0.96186537,  0.8166533 ,  0.54394835, -0.627869  ,
-0.726707  , -0.5601597 , -0.14287454,  0.0786843 , -0.19265983,
 0.655331  ,  0.69784737,  0.32879114,  0.5070215 , -0.48422274,
 0.85959285, -0.76183695, -0.8617145 , -0.5963389 , -0.09587274,
-0.93534106,  0.66678095,  0.35500234,  0.5884894 , -0.43380925,
-0.5236545 , -0.787733  ,  0.60011965,  0.19763757,  0.9121581 ,
-0.22849797, -0.6094818 , -0.50804675, -0.7245058 , -0.12123018,
-0.0367312 , -0.14356132, -0.12071957, -0.7052898 ,  0.44657195,
 0.32162455,  0.4233121 , -0.51390326,  0.96877027,  0.9999993 ,
 0.83680826,  0.6068135 ,  0.48950708, -0.9995693 , -0.8199097 ,
```

```
 0.9999737 , -0.9688966 , -1.        , -0.79203635, -0.38465968,
 0.15010926, -1.        , -0.07409794,  0.11116396, -0.5812642 ,
 0.46954635,  0.88659203,  0.78593606, -1.        ,  0.6759215 ,
 0.85693985, -0.69559586,  0.71892434, -0.35184866,  0.8558557 ,
 0.02280325,  0.5379513 , -0.18867256,  0.20705144, -0.6795617 ,
-0.6724148 , -0.3723054 , -0.5700406 ,  0.9982937 ,  0.09954868,
-0.85268897, -0.61288214,  0.54186434, -0.02283256, -0.3802247 ,
-0.8514071 , -0.20104828,  0.18678845,  0.50533307,  0.24392857,
 0.23983045, -0.36754268,  0.27392432,  0.14140348, -0.00668742,
 0.6949476 , -0.7999368 , -0.18021321, -0.25588554, -0.05541642,
-0.29810444, -0.92562276,  0.81629956, -0.32016778,  0.67978835,
 1.        ,  0.5250536 , -0.6078582 ,  0.50854594,  0.253711  ,
 0.03503525,  1.        ,  0.8028996 , -0.88136345, -0.6841441 ,
 0.67018145, -0.4302866 , -0.62930036,  0.99667823, -0.3564602 ,
-0.4152862 , -0.01637087,  0.8919304 , -0.9444473 ,  0.990642  ,
-0.6202186 , -0.8652292 ,  0.8398232 ,  0.6462517 , -0.26379654,
-0.6831857 ,  0.07372912, -0.5444585 ,  0.29530516, -0.6811525 ,
 0.60219914,  0.1339953 , -0.06221671,  0.6656015 , -0.39688805,
-0.6749006 ,  0.34520262, -0.60926527, -0.14574575,  0.8665897 ,
 0.44202596, -0.17941171,  0.1370639 , -0.13074614, -0.24475507,
-0.90795743,  0.567175  ,  1.        , -0.12939744,  0.73375976,
-0.09851032, -0.0819903 , -0.06798995,  0.36033723,  0.4608149 ,
-0.16137794, -0.6413537 ,  0.36447695, -0.4019046 , -0.94279385,
 0.33668894,  0.10700357, -0.08686241,  0.99982786,  0.19443935,
 0.26778868,  0.03445277,  0.92345726,  0.05086514,  0.31566083,
 0.50754637,  0.9228061 , -0.11244283,  0.7093428 ,  0.43788946,
-0.53869873, -0.1553493 , -0.5682935 ,  0.03738712, -0.83025855,
 0.04560403, -0.8314194 ,  0.8910811 ,  0.9079797 ,  0.33861703,
 0.10296188,  0.6555968 ,  1.        , -0.88848203,  0.16696753,
 0.5622083 , -0.05940764, -0.9993758 , -0.16683872, -0.19263221,
 0.08520819, -0.57204276, -0.26912558,  0.08731644, -0.8726315 ,
 0.5479681 ,  0.39123225, -0.75649625, -0.8971171 , -0.27833933,
 0.57108265, -0.02563838, -0.97662634, -0.52592283, -0.26335913,
 0.03208682, -0.22296266, -0.6944577 ,  0.02822472, -0.28071588,
 0.39100036, -0.14882833,  0.6052244 ,  0.63480484,  0.8061493 ,
-0.8523697 , -0.42808175, -0.05561546, -0.63380444,  0.52067006,
-0.7657652 , -0.8439526 , -0.17108092,  1.        , -0.05442048,
 0.60706365,  0.4992314 ,  0.3690213 , -0.16289152,  0.24505043,
 0.92528665,  0.22433391, -0.30461547, -0.50231165,  0.411674  ,
-0.45100442,  0.52243316,  0.41183886,  0.60915154,  0.2549151 ,
 0.71456313,  0.20587085, -0.03063838,  0.0957486 ,  0.9572757 ,
-0.16835989, -0.11695716, -0.39266267, -0.10726684, -0.2874045 ,
 0.33275425,  1.        ,  0.30811134,  0.65996206, -0.9377766 ,
-0.54616386, -0.7871911 ,  0.99999785,  0.6207103 , -0.44167328,
 0.5145529 ,  0.5095189 , -0.19141568,  0.33160412, -0.18261477,
-0.2919685 ,  0.27162957,  0.14202197,  0.7772855 , -0.3772172 ,
-0.880914  , -0.60588545,  0.28090662, -0.81398517,  0.99981594,
```

```
-0.53860176, -0.25904024, -0.1646331 , -0.5620859 , -0.5492251 ,
 0.09809349, -0.8571454 , -0.16794421,  0.00127427,  0.7835233 ,
 0.24669304, -0.65947783, -0.8542363 ,  0.77735305,  0.39669505,
-0.6388964 , -0.7680781 ,  0.7989103 , -0.8719203 ,  0.4707688 ,
 0.99999905,  0.35008925,  0.20026024, -0.00357179, -0.14283592,
 0.23867753, -0.5446804 ,  0.3214971 , -0.79477596, -0.27013722,
-0.24085617,  0.34835425, -0.14272007, -0.8253409 ,  0.257251  ,
 0.20217855, -0.6259333 , -0.55393386,  0.03889284,  0.3645683 ,
 0.6358549 , -0.23581143, -0.08056157,  0.11550374,  0.02292791,
-0.5434361 , -0.26907337, -0.3817061 , -0.99997556,  0.50361145,
-1.        ,  0.44884732,  0.01614935, -0.14968823,  0.6509744 ,
 0.5061133 ,  0.49569085, -0.23077025, -0.700562  ,  0.66291225,
 0.43474787, -0.31810153, -0.3049655 , -0.45363346,  0.21092984,
-0.01006573,  0.21341722, -0.67999095,  0.665682  , -0.26421982,
 1.        ,  0.1455384 , -0.34065378, -0.5552133 ,  0.2517574 ,
-0.20953272,  0.99999964, -0.4640369 , -0.86137503,  0.31059983,
-0.79067117, -0.683044  ,  0.37750393,  0.00977041, -0.7032061 ,
-0.7832784 ,  0.6932373 ,  0.59866244, -0.6551872 ,  0.408208  ,
-0.33261126, -0.5451027 ,  0.1286118 ,  0.8130428 ,  0.9411659 ,
 0.2654561 ,  0.6465997 , -0.47814903,  0.08139641,  0.81585103,
 0.26410633, -0.02503392,  0.22744216,  1.        ,  0.45141053,
-0.78067595,  0.4204063 , -0.8349061 , -0.16405927, -0.87020516,
 0.24746135,  0.22271064,  0.7679804 , -0.27900136,  0.7890243 ,
-0.70328116,  0.03303251, -0.40275344, -0.32862028,  0.2905226 ,
-0.76074237, -0.8972959 , -0.8873006 ,  0.6180563 , -0.30493283,
-0.07593845,  0.27093002,  0.1498122 ,  0.3805027 ,  0.41765144,
-1.        ,  0.803459  ,  0.29456514,  0.60832375,  0.80494756,
 0.6780596 ,  0.58588487,  0.32866186, -0.90430677, -0.35899773,
-0.195941  , -0.20472781,  0.64030415,  0.7528085 ,  0.57756484,
 0.19472875, -0.43566114, -0.4729009 , -0.2564997 , -0.8012523 ,
-0.96441597,  0.40390006, -0.36510128, -0.56354636,  0.88209265,
-0.37550512, -0.07116201,  0.02441978, -0.8021945 ,  0.65386176,
 0.62444806, -0.01478089,  0.00386956,  0.44045782,  0.55791   ,
 0.79091877,  0.9234506 , -0.77168334,  0.653177  , -0.67965513,
 0.2937751 ,  0.81573653, -0.87198037,  0.12073833,  0.5433519 ,
 0.05705976,  0.20133454, -0.15797368, -0.6467374 ,  0.77185166,
-0.10124951,  0.53214806, -0.31655902,  0.08648318, -0.47987378,
-0.23402078, -0.5035185 , -0.5291463 ,  0.63656276, -0.08323704,
 0.644091  ,  0.7778061 , -0.10646352, -0.555375  , -0.17449316,
-0.47210306, -0.73606586,  0.47677922,  0.11129446, -0.12331934,
 0.6223576 , -0.10882687,  0.90658367,  0.00600077, -0.34572726,
-0.15421392, -0.46045142,  0.5624086 , -0.6612781 , -0.5444964 ,
-0.49284026,  0.6287911 ,  0.2729571 ,  0.9999757 , -0.3444969 ,
-0.5707927 , -0.62781453, -0.31368667,  0.30566216, -0.45169756,
-1.        ,  0.2816248 , -0.70476335,  0.4582703 , -0.31213644,
 0.480618  , -0.6419509 , -0.81418085, -0.20872763,  0.50627804,
 0.66102636, -0.4779182 , -0.54610217,  0.67544883, -0.6030795 ,
```

```
         0.9261357 ,  0.61045146, -0.37918216,  0.30536515,  0.715992  ,
        -0.6844849 , -0.48579797,  0.5437323 ]], dtype=float32)>
```

[ ]:
```
bert_encoder = hub.KerasLayer("https://tfhub.dev/tensorflow/
  ↪bert_en_uncased_L-12_H-768_A-12/4")
```

[ ]:
```
# Bert layers
text_input = tf.keras.layers.Input(shape=(), dtype=tf.string, name='text')
preprocessed_text = bert_preprocess(text_input)
outputs = bert_encoder(preprocessed_text)

# Neural network layers
l = tf.keras.layers.Dropout(0.1, name="dropout")(outputs['pooled_output'])
l = tf.keras.layers.Dense(1, activation='sigmoid', name="output")(l)

# Use inputs and outputs to construct a final model
model = tf.keras.Model(inputs=[text_input], outputs = [l])
```

[ ]:
```
model.summary()
```

```
Model: "model"
__________________________________________________________________________________
__________________
 Layer (type)                Output Shape              Param #   Connected to
==================================================================================
==================
 text (InputLayer)           [(None,)]                 0         []

 keras_layer (KerasLayer)    {'input_type_ids': (None,    0
['text[0][0]']
                             128),
                              'input_mask': (None, 128)
                             , 'input_word_ids': (None,
                              128)}

__________________________________________________________________________________
__________________
 Layer (type)                Output Shape              Param #   Connected to
==================================================================================
==================
 text (InputLayer)           [(None,)]                 0         []

 keras_layer (KerasLayer)    {'input_type_ids': (None,    0
['text[0][0]']
                             128),
                              'input_mask': (None, 128)
                             , 'input_word_ids': (None,
                              128)}
```

```
 keras_layer_2 (KerasLayer)  {'pooled_output': (None, 7   1094822
['keras_layer[0][0]',
                             68),                          41
'keras_layer[0][1]',
                              'sequence_output': (None,
'keras_layer[0][2]']
                              128, 768),
                              'encoder_outputs': [(None
                             , 128, 768),
                              (None, 128, 768),
                              (None, 128, 768),
                              (None, 128, 768),
                              (None, 128, 768),
                              (None, 128, 768),
                              (None, 128, 768),
                              (None, 128, 768),
                              (None, 128, 768),
                              (None, 128, 768),
                              (None, 128, 768),
                              (None, 128, 768)],
                              'default': (None, 768)}

 dropout (Dropout)           (None, 768)                   0
['keras_layer_2[0][13]']

 output (Dense)              (None, 1)                     769
['dropout[0][0]']

==================================================================================
==================
Total params: 109483010 (417.64 MB)
Trainable params: 769 (3.00 KB)
Non-trainable params: 109482241 (417.64 MB)
----------------------------------------------------------------------------------
------------------
```

[ ]:
```
METRICS = [
      tf.keras.metrics.BinaryAccuracy(name='accuracy'),
      tf.keras.metrics.Precision(name='precision'),
      tf.keras.metrics.Recall(name='recall')
]

model.compile(optimizer='adam',
              loss='binary_crossentropy',
              metrics=METRICS)
```

0.4.1 Model-Train

[ ]:
```
model.fit(x_train, y_train, epochs=1)
```

```
147/147 [==============================] - 992s 7s/step - loss: -3.5112 -
accuracy: 0.5363 - precision: 0.8532 - recall: 0.9985
```

[ ]:
```
<keras.src.callbacks.History at 0x7f994bc31010>
```

[ ]:
```
model.evaluate(x_test, y_test)
```

```
37/37 [==============================] - 9497s 264s/step - loss: -6.9543 -
accuracy: 0.5304 - precision: 0.8506 - recall: 0.9970
```

[ ]:
```
[-6.9542694091796875,
 0.5303678512573242,
 0.8506437540054321,
 0.9969819188117981]
```

[ ]:
```
y_predicted = model.predict(x_test)
y_predicted = y_predicted.flatten()
```

```
37/37 [==============================] - 260s 7s/step
```

0.4.2 Model - Evaluate

[ ]:
```
from sklearn.metrics import confusion_matrix, classification_report
from matplotlib import pyplot as plt
import seaborn as sn
```

[ ]:
```
y_predicted = np.where(y_predicted > 0.5, 1, 0)
y_predicted
```

[ ]:
```
array([1, 1, 1, …, 1, 1, 1])
```

[ ]:
```
cm = confusion_matrix(y_test, y_predicted)
cm
```

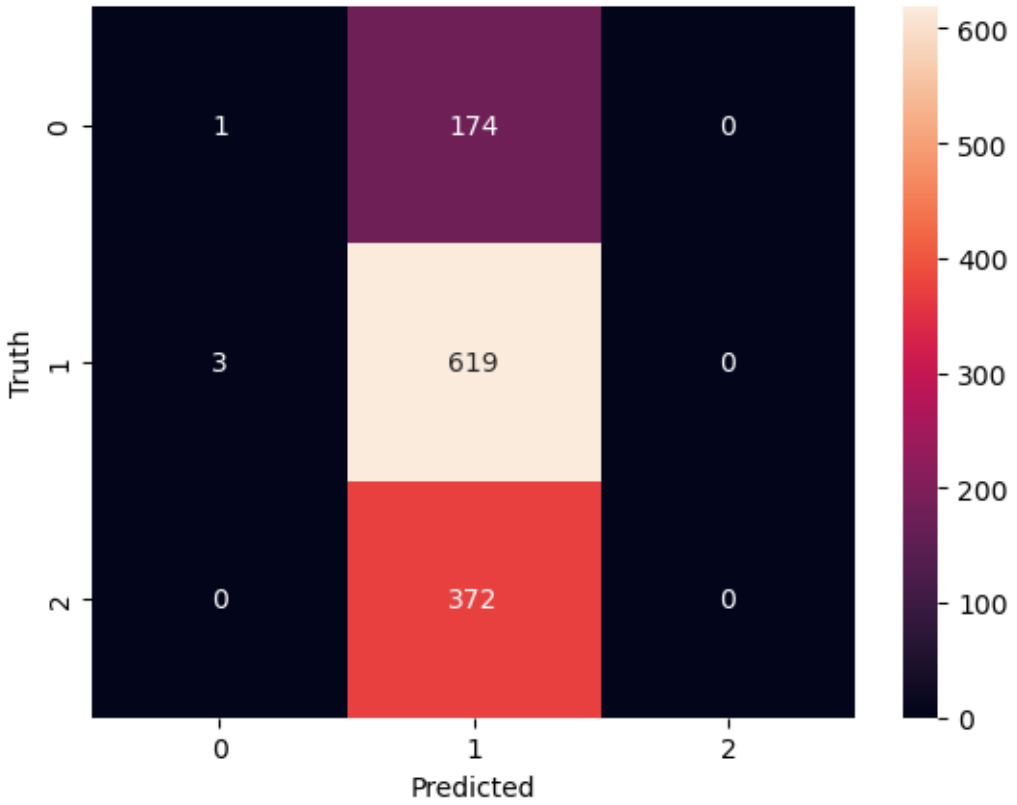
```
/home/chocomenta/anaconda3/lib/python3.11/site-
packages/sklearn/utils/validation.py:605: FutureWarning: is_sparse is deprecated
and will be removed in a future version. Check `isinstance(dtype,
pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
/home/chocomenta/anaconda3/lib/python3.11/site-
packages/sklearn/utils/validation.py:614: FutureWarning: is_sparse is deprecated
and will be removed in a future version. Check `isinstance(dtype,
pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
/home/chocomenta/anaconda3/lib/python3.11/site-
```

```
packages/sklearn/utils/validation.py:605: FutureWarning: is_sparse is deprecated
and will be removed in a future version. Check `isinstance(dtype,
pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
/home/chocomenta/anaconda3/lib/python3.11/site-
packages/sklearn/utils/validation.py:614: FutureWarning: is_sparse is deprecated
and will be removed in a future version. Check `isinstance(dtype,
pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
/home/chocomenta/anaconda3/lib/python3.11/site-
packages/sklearn/utils/validation.py:605: FutureWarning: is_sparse is deprecated
and will be removed in a future version. Check `isinstance(dtype,
pd.SparseDtype)` instead.
  if is_sparse(pd_dtype):
/home/chocomenta/anaconda3/lib/python3.11/site-
packages/sklearn/utils/validation.py:614: FutureWarning: is_sparse is deprecated
and will be removed in a future version. Check `isinstance(dtype,
pd.SparseDtype)` instead.
  if is_sparse(pd_dtype) or not is_extension_array_dtype(pd_dtype):
```

[ ]:
```
array([[  1, 174,   0],
       [  3, 619,   0],
       [  0, 372,   0]])
```

[ ]:
```
sn.heatmap(cm, annot=True, fmt='d')
plt.xlabel('Predicted')
plt.ylabel('Truth')
```

[ ]:
```
Text(50.722222222222214, 0.5, 'Truth')
```


1
174
0
3
619
0
0
372
0
Truth
0
1
2
0
1
2
Predicted
600
500
400
300
200
100
0