

Predicting_Heart_Problem_with_BERT_in_Tensorflow RAW 5

July 30, 2020

##Mounting Google Drive

```
[1]: from google.colab import drive
drive.mount("/GD")
```

Go to this URL in a browser: https://accounts.google.com/o/oauth2/auth?client_id=947318989803-6bn6qk8qdgf4n4g3pfee6491hc0brc4i.apps.googleusercontent.com&redirect_uri=urn%3aietf%3awg%3aoauth%3a2.0%3aob&response_type=code&scope=email%20https%3a%2f%2fwww.googleapis.com%2fauth%2fdocs.test%20https%3a%2f%2fwww.googleapis.com%2fauth%2fdrive%20https%3a%2f%2fwww.googleapis.com%2fauth%2fdrive.photos.readonly%20https%3a%2f%2fwww.googleapis.com%2fauth%2fpeopleapi.readonly

Enter your authorization code:

.....

Mounted at /GD

0.1 Importing Necessary Libraries

```
[ ]: !pip install tensorflow==1.15.0
import tensorflow as tf
print(tf.__version__)
```

```
[1]: import pandas as pd
import tensorflow as tf
import tensorflow_hub as hub
from datetime import datetime
from sklearn.model_selection import train_test_split
import os

print("tensorflow version : ", tf.__version__)
print("tensorflow_hub version : ", hub.__version__)
```

tensorflow version : 1.15.0
tensorflow_hub version : 0.8.0

```
[2]: #Installing BERT module
!pip install bert-tensorflow
```

Requirement already satisfied: bert-tensorflow in /usr/local/lib/python3.6/dist-packages (1.0.1)

Requirement already satisfied: six in /usr/local/lib/python3.6/dist-packages (from bert-tensorflow) (1.15.0)

```
[3]: #Importing BERT modules
import bert
from bert import run_classifier
from bert import optimization
from bert import tokenization
```

WARNING:tensorflow:From /usr/local/lib/python3.6/dist-packages/bert/optimization.py:87: The name tf.train.Optimizer is deprecated. Please use tf.compat.v1.train.Optimizer instead.

0.2 ##Setting The Output Directory

While fine-tuning the model, we will save the training checkpoints and the model in an output directory so that we can use the trained model for our predictions later.

The following code block sets an output directory :

```
[4]: # Set the output directory for saving model file
OUTPUT_DIR = '/GD/My Drive/Colab Notebooks/5epochs'

#@markdown Whether or not to clear/delete the directory and create a new one
DO_DELETE = False #@param {type:"boolean"}

if DO_DELETE:
    try:
        tf.gfile.DeleteRecursively(OUTPUT_DIR)
    except:
        pass

tf.gfile.MakeDirs(OUTPUT_DIR)
print('***** Model output directory: {} *****'.format(OUTPUT_DIR))
```

***** Model output directory: /GD/My Drive/Colab Notebooks/5epochs *****

0.3 ##Loading The Data

We will now load the data from a Google Drive directory and will also split the training set in to training and validation sets.

```
[5]: train = pd.read_csv("/GD/My Drive/Colab Notebooks/5epochs/raw_sampled_data.
    ↪ csv", encoding = "ISO-8859-1")
train['question'] = train['question'].apply(str)
```

```
from sklearn.model_selection import train_test_split

train, val = train_test_split(train, test_size = 0.2, random_state = 100)
```

```
[6]: #Training set sample
train.head(15)
```

```
[6]:
```

	category	...	question
629	hipertensi	...	Assalamualaikum dok, sy mau bertanya. Ibu saya...
669	hipertensi	...	Dok, 1 bulan yang lalu saya periksa ke puskesmas...
25	aritmia	...	Dok saya wanita berusia 24 tahu. Sejak 5 tahun...
88	aritmia	...	Dok dada trasa aga sesak tapi tidak batuk, 3ha...
395	gagal-jantung	...	Halo dok, sy pengidap jantung bocor, selalu ko...
685	hipertensi	...	Permisi, Dok. saya berusia 21th, 3minggu yang ...
1049	stroke	...	Dok, 6 bulan lalu saya terkena stroke (pecah p...
975	serangan-jantung	...	Dok saya kaya terasa sesak dok terus kadang da...
356	gagal-jantung	...	Bapak saya skrg lg d rawat di RS\r\nkepala pus...
146	aritmia	...	Dok saya mau bertanya? Kenapa denyut nadi bisa...
1281	stroke	...	Selamat siang dokter, Maaf saya ingin bertanya,...
45	aritmia	...	Maaf Dok, detakan jantung saya tiba-tiba kuat ...
330	gagal-jantung	...	Pagi dok , saya mau tanya dok mamah saya tanga...
160	aritmia	...	Dok, akhir" ini ketika saya menaiki tangga jan...
306	gagal-jantung	...	dok,, ibu saya kira" umur 37th.. saring kaget ...

[15 rows x 6 columns]

```
[7]: print("Training Set Shape :", train.shape)
print("Validation Set Shape :", val.shape)
#print("Test Set Shape :", test.shape)
```

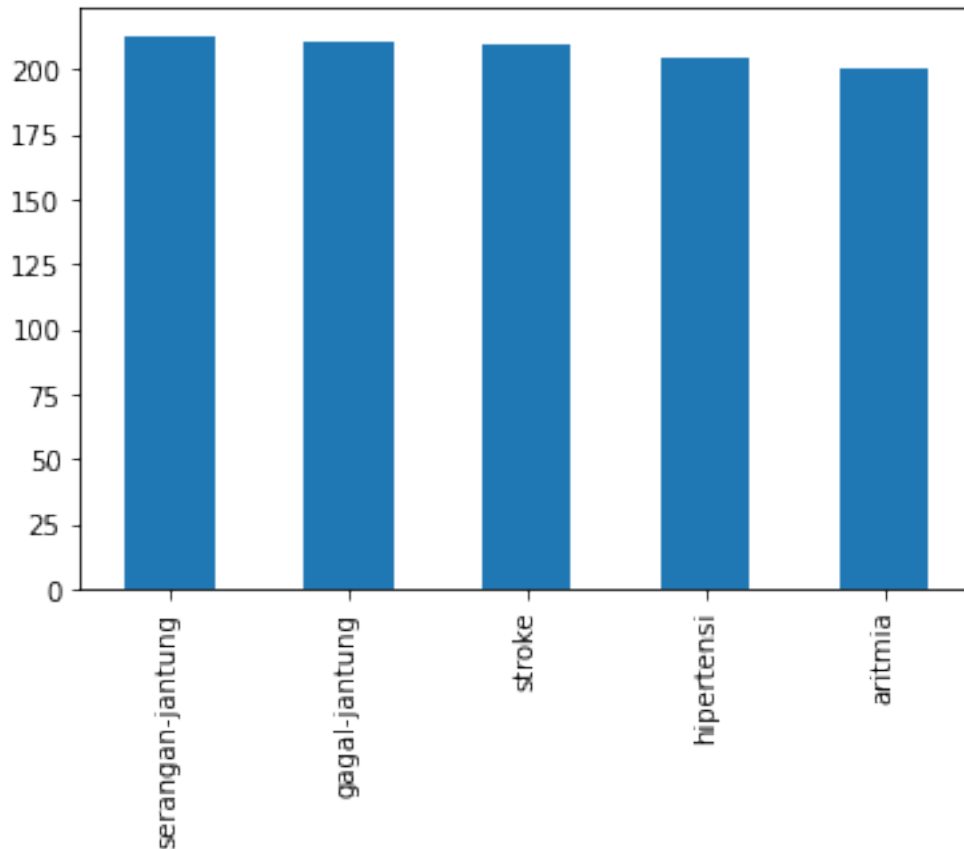
```
Training Set Shape : (1040, 6)
Validation Set Shape : (260, 6)
```

```
[8]: #unique classes
train['category'].unique()
```

```
[8]: array(['hipertensi', 'aritmia', 'gagal-jantung', 'stroke',
        'serangan-jantung'], dtype=object)
```

```
[9]: #Distribution of classes
train['category'].value_counts().plot(kind = 'bar')
```

```
[9]: <matplotlib.axes._subplots.AxesSubplot at 0x7faebbf15f60>
```



```
[10]: DATA_COLUMN = 'question'
      LABEL_COLUMN = 'category'
      # The list containing all the classes (train['SECTION'].unique())
      label_list = list(train['category'].unique())
```

0.4 Data Preprocessing

BERT model accept only a specific type of input and the datasets are usually structured to have the following four features:

- `guid` : A unique id that represents an observation.
- `text_a` : The text we need to classify into given categories
- `text_b`: It is used when we're training a model to understand the relationship between sentences and it does not apply for classification problems.
- `label`: It consists of the labels or classes or categories that a given text belongs to.

In our dataset we have `text_a` and `label`. The following code block will create objects for each of the above mentioned features for all the records in our dataset using the `InputExample` class provided in the BERT library.

```
[11]: train_InputExamples = train.apply(lambda x: bert.run_classifier.
      ↳ InputExample(guid=None,
                                                    text_a = "
      ↳ x[DATA_COLUMN],
                                                    text_b = "
      ↳ None,
                                                    label = "
      ↳ x[LABEL_COLUMN]), axis = 1)

val_InputExamples = val.apply(lambda x: bert.run_classifier.
      ↳ InputExample(guid=None,
                                                    text_a = "
      ↳ x[DATA_COLUMN],
                                                    text_b = "
      ↳ None,
                                                    label = "
      ↳ x[LABEL_COLUMN]), axis = 1)
```

```
[12]: train_InputExamples
```

```
[12]: 629    <bert.run_classifier.InputExample object at 0x...
      669    <bert.run_classifier.InputExample object at 0x...
      25     <bert.run_classifier.InputExample object at 0x...
      88     <bert.run_classifier.InputExample object at 0x...
      395    <bert.run_classifier.InputExample object at 0x...
      ...
      802    <bert.run_classifier.InputExample object at 0x...
      53     <bert.run_classifier.InputExample object at 0x...
      350    <bert.run_classifier.InputExample object at 0x...
      79     <bert.run_classifier.InputExample object at 0x...
      792    <bert.run_classifier.InputExample object at 0x...
      Length: 1040, dtype: object
```

```
[13]: print("Row 0 - guid of training set : ", train_InputExamples.iloc[0].guid)
      print("\n_____ \nRow 0 - text_a of training set : ", train_InputExamples.
      ↳ iloc[0].text_a)
      print("\n_____ \nRow 0 - text_b of training set : ", train_InputExamples.
      ↳ iloc[0].text_b)
      print("\n_____ \nRow 0 - label of training set : ", train_InputExamples.
      ↳ iloc[0].label)
```

Row 0 - guid of training set : None

 Row 0 - text_a of training set : Assalamualaikum dok, sy mau bertanya. Ibu saya menderita hipertensi. Sudah di bawa ke dokter dan diberi obat penurun hipertensi. Tapi kenapa kok tekanan darahnya tidak kunjung normal? Padahal sudah

rutin minum obat & mengkonsumsi banyak buah2an (timun, semangka, dll). Malah badannya terasa lemas, pusing & penglihatan kabur. Terimakasih. Wassalamualaikum
õ õ

Row 0 - text_b of training set : None

Row 0 - label of training set : hipertensi

We will now get down to business with the pretrained BERT. In this example we will use the bert_uncased_L-12_H-768_A-12/1 model. To check all available versions click [here](#).

We will be using the vocab.txt file in the model to map the words in the dataset to indexes. Also the loaded BERT model is trained on uncased/lowercase data and hence the data we feed to train the model should also be of lowercase.

The following code block loads the pre-trained BERT model and initializes a tokenizer object for tokenizing the texts.

```
[14]: # This is a path to an uncased (all lowercase) version of BERT
BERT_MODEL_HUB = "https://tfhub.dev/google/bert_multi_cased_L-12_H-768_A-12/1"

def create_tokenizer_from_hub_module():
    """Get the vocab file and casing info from the Hub module."""
    with tf.Graph().as_default():
        bert_module = hub.Module(BERT_MODEL_HUB)
        tokenization_info = bert_module(signature="tokenization_info", as_dict=True)
        with tf.Session() as sess:
            vocab_file, do_lower_case = sess.run([tokenization_info["vocab_file"],
                                                  tokenization_info["do_lower_case"]])

    return bert.tokenization.FullTokenizer(
        vocab_file=vocab_file, do_lower_case=do_lower_case)

tokenizer = create_tokenizer_from_hub_module()
```

INFO:tensorflow:Saver not created because there are no variables in the graph to restore

INFO:tensorflow:Saver not created because there are no variables in the graph to restore

WARNING:tensorflow:From /usr/local/lib/python3.6/dist-packages/bert/tokenization.py:125: The name tf.gfile.GFile is deprecated. Please use tf.io.gfile.GFile instead.

WARNING:tensorflow:From /usr/local/lib/python3.6/dist-

packages/bert/tokenization.py:125: The name tf.gfile.GFile is deprecated. Please use tf.io.gfile.GFile instead.

```
[15]: #Here is what the tokenised sample of the first training set observation looks  
      ↳ like  
      print(tokenizer.tokenize(train_InputExamples.iloc[9].text_a))
```

```
['Dok', 'saya', 'mau', 'bertan', '##ya', '?', 'Ken', '##apa', 'den', '##yu',  
 '##t', 'nad', '##i', 'bisa', 'di', 'pak', '##ai', 'untuk', 'menentukan', 'be',  
 '##ban', 'kerja', 'fi', '##sik', ',', 'tingkat', 'kes', '##eh', '##atan', ',',  
 'tingkat', 'ke', '##bu', '##garan', 'dan', 'tingkat', 'stress', '?']
```

We will now format out text in to input features which the BERT model expects. We will also set a sequence length which will be the length of the input features.

```
[16]: max_len = max([len(tokenizer.tokenize(train_InputExamples.iloc[IDX].text_a))  
      ↳ for IDX in range(1040)])  
      print('Max length: ', max_len)
```

Max length: 2161

```
[17]: # We'll set sequences to be at most 128 tokens long.  
      MAX_SEQ_LENGTH = 256  
  
      # Convert our train and validation features to InputFeatures that BERT  
      ↳ understands.  
      train_features = bert.run_classifier.  
      ↳ convert_examples_to_features(train_InputExamples, label_list,  
      ↳ MAX_SEQ_LENGTH, tokenizer)  
  
      val_features = bert.run_classifier.  
      ↳ convert_examples_to_features(val_InputExamples, label_list, MAX_SEQ_LENGTH,  
      ↳ tokenizer)
```

WARNING:tensorflow:From /usr/local/lib/python3.6/dist-packages/bert/run_classifier.py:774: The name tf.logging.info is deprecated. Please use tf.compat.v1.logging.info instead.

WARNING:tensorflow:From /usr/local/lib/python3.6/dist-packages/bert/run_classifier.py:774: The name tf.logging.info is deprecated. Please use tf.compat.v1.logging.info instead.

INFO:tensorflow:Writing example 0 of 1040

INFO:tensorflow:Writing example 0 of 1040

INFO:tensorflow:*** Example ***

INFO:tensorflow:*** Example ***

INFO:tensorflow:guid: None

INFO:tensorflow:guid: None

INFO:tensorflow:tokens: [CLS] Ass ##ala ##mua ##lai ##kum dok , sy mau bertan
##ya . Ibu saya men ##der ##ita hip ##erten ##si . Sud ##ah di ba ##wa ke dokter
dan diberi oba ##t pen ##uru ##n hip ##erten ##si . Ta ##pi ken ##apa ko ##k
tekanan darah ##nya tidak kun ##jung normal ? Pada ##hal sudah ru ##tin min ##um
oba ##t & men ##g ##kon ##sum ##si banyak buah ##2 ##an (tim ##un , sem ##ang
##ka , dl ##l) . Mala ##h badan ##nya ter ##asa lema ##s , pus ##ing & pen
##gli ##hat ##an ka ##bur . Ter ##ima ##kasi ##h . Was ##sala ##mua ##lai ##kum
ð ##ð [SEP]

INFO:tensorflow:tokens: [CLS] Ass ##ala ##mua ##lai ##kum dok , sy mau bertan
##ya . Ibu saya men ##der ##ita hip ##erten ##si . Sud ##ah di ba ##wa ke dokter
dan diberi oba ##t pen ##uru ##n hip ##erten ##si . Ta ##pi ken ##apa ko ##k
tekanan darah ##nya tidak kun ##jung normal ? Pada ##hal sudah ru ##tin min ##um
oba ##t & men ##g ##kon ##sum ##si banyak buah ##2 ##an (tim ##un , sem ##ang
##ka , dl ##l) . Mala ##h badan ##nya ter ##asa lema ##s , pus ##ing & pen
##gli ##hat ##an ka ##bur . Ter ##ima ##kasi ##h . Was ##sala ##mua ##lai ##kum
ð ##ð [SEP]

INFO:tensorflow:input_ids: 101 77014 13322 78314 31181 36811 17674 117 12261
43024 49111 10679 119 68096 64981 10588 11304 11622 25377 26645 10449 119 13352
12257 10120 15688 11037 11163 96140 10215 50479 35355 10123 66558 25279 10115
25377 26645 10449 119 14248 12675 67680 46757 11252 10174 93131 43947 10676
11868 13158 30425 16626 136 12270 18453 25147 13483 15364 13484 10465 35355
10123 111 10588 10240 17423 31417 10449 15175 21988 10729 10206 113 19604 11107
117 11531 11889 10371 117 63940 10161 114 119 58335 10237 51463 10676 12718
23031 93661 10107 117 46960 10230 111 66558 20986 19180 10206 10730 34660 119
65272 12443 37997 10237 119 22034 104655 78314 31181 36811 270 12332 102 0 0 0 0
0
0
0
0 0

INFO:tensorflow:input_ids: 101 77014 13322 78314 31181 36811 17674 117 12261
43024 49111 10679 119 68096 64981 10588 11304 11622 25377 26645 10449 119 13352
12257 10120 15688 11037 11163 96140 10215 50479 35355 10123 66558 25279 10115
25377 26645 10449 119 14248 12675 67680 46757 11252 10174 93131 43947 10676
11868 13158 30425 16626 136 12270 18453 25147 13483 15364 13484 10465 35355
10123 111 10588 10240 17423 31417 10449 15175 21988 10729 10206 113 19604 11107
117 11531 11889 10371 117 63940 10161 114 119 58335 10237 51463 10676 12718
23031 93661 10107 117 46960 10230 111 66558 20986 19180 10206 10730 34660 119
65272 12443 37997 10237 119 22034 104655 78314 31181 36811 270 12332 102 0 0 0 0
0
0
0
0 0

[illegible][illegible][illegible][illegible]

```
INFO:tensorflow:label: hipertensi (id = 0)
```

```
INFO:tensorflow:label: hipertensi (id = 0)
```

```
INFO:tensorflow:*** Example ***
```

INFO:tensorflow:*** Example ***

INFO:tensorflow:guid: None

INFO:tensorflow:guid: None

INFO:tensorflow:tokens: [CLS] Dok , 1 bulan yang lalu saya per ##iks ##a ke pus
##kes ##mas dan ten ##si saya 180 / 120 dan diberikan oba ##t cap ##top ##ril 25
##m ##g dan corsa ##neur ##on . du ##lu pernah diberikan kom ##bina ##si cap
##top ##ril dan fur ##ose ##mide . yang saya tany ##akan lebih baik mana dari 2
kom ##bina ##si oba ##t tersebut ? [SEP]

INFO:tensorflow:tokens: [CLS] Dok , 1 bulan yang lalu saya per ##iks ##a ke pus
##kes ##mas dan ten ##si saya 180 / 120 dan diberikan oba ##t cap ##top ##ril 25
##m ##g dan corsa ##neur ##on . du ##lu pernah diberikan kom ##bina ##si cap

##top ##ril dan fur ##ose ##mide . yang saya tany ##akan lebih baik mana dari 2
kom ##bina ##si oba ##t tersebut ? [SEP]

INFO:tensorflow:input_ids: 101 85566 117 122 12385 10265 31288 64981 10178 40670
10113 11163 46960 21885 12922 10215 11769 10449 64981 13912 120 12048 10215
44141 35355 10123 13337 37253 31860 10258 10147 10240 10215 50583 38780 10263
119 10168 11435 21407 44141 12240 29368 10449 13337 37253 31860 10215 61001
14569 58137 119 10265 64981 89875 27125 13394 24604 18970 10397 123 12240 29368
10449 35355 10123 12848 136 102 0
0
0
0
0
0 0 0
0 0 0

INFO:tensorflow:input_ids: 101 85566 117 122 12385 10265 31288 64981 10178 40670
10113 11163 46960 21885 12922 10215 11769 10449 64981 13912 120 12048 10215
44141 35355 10123 13337 37253 31860 10258 10147 10240 10215 50583 38780 10263
119 10168 11435 21407 44141 12240 29368 10449 13337 37253 31860 10215 61001
14569 58137 119 10265 64981 89875 27125 13394 24604 18970 10397 123 12240 29368
10449 35355 10123 12848 136 102 0
0
0
0
0
0 0 0
0 0 0

INFO:tensorflow:input_mask: 1
1
1 1 1 0
0
0
0
0
0 0

INFO:tensorflow:input_mask: 1
1
1 1 1 0
0
0
0
0
0 0

INFO:tensorflow:segment_ids: 0
0
0
0
0
0
0 0

##mang sy men ##gida ##p hip ##erten ##si ? A ##paka ##h ada kor ##elas ##i d
##gn ke ##bo ##cora ##n jan ##tung dan tekanan darah tinggi ? [SEP]

INFO:tensorflow:input_ids: 101 67679 17674 117 12261 66558 32556 10410 63923
23091 20506 49167 117 56894 55605 24590 123 12385 46233 111 10794 10174 93131
43947 117 111 31102 10676 16626 113 12718 76010 10376 25377 44073 10449 114 117
46233 10676 23057 178 10240 92287 10237 70733 11471 17502 119 14248 12675 67680
46757 46687 12261 56894 66558 10230 24590 10126 100025 10291 76306 75902 12438
27300 136 138 103208 10237 10592 176 10676 10198 63952 10116 11754 10911 45306
12261 10588 32556 10410 25377 26645 10449 136 138 103208 10237 15290 33705 50567
10116 172 19962 11163 11790 75347 10115 63923 23091 10215 93131 43947 23057 136
102 0
0
0
0 0

INFO:tensorflow:input_ids: 101 67679 17674 117 12261 66558 32556 10410 63923
23091 20506 49167 117 56894 55605 24590 123 12385 46233 111 10794 10174 93131
43947 117 111 31102 10676 16626 113 12718 76010 10376 25377 44073 10449 114 117
46233 10676 23057 178 10240 92287 10237 70733 11471 17502 119 14248 12675 67680
46757 46687 12261 56894 66558 10230 24590 10126 100025 10291 76306 75902 12438
27300 136 138 103208 10237 10592 176 10676 10198 63952 10116 11754 10911 45306
12261 10588 32556 10410 25377 26645 10449 136 138 103208 10237 15290 33705 50567
10116 172 19962 11163 11790 75347 10115 63923 23091 10215 93131 43947 23057 136
102 0
0
0
0 0

INFO:tensorflow:input_mask: 1
1
1
1 1 0
0
0
0 0

INFO:tensorflow:input_mask: 1
1
1
1 1 0
0
0
0 0

INFO:tensorflow:segment_ids: 0
0
0
0
0 0


```
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

[illegible]

```
INFO:tensorflow:label: stroke (id = 3)
```

```
INFO:tensorflow:*** Example ***
```

INFO:tensorflow:guid: None

INFO:tensorflow:tokens: [CLS] Selama ##t sia ##ng dok . Say ##a ingin bertan
##ya . A ##paka ##h orang y ##g mengalami sakit jan ##tung pada stadium 4 bisa
be ##rku ##rang menjadi stadium 3 , stadium 3 menjadi stadium 2 , stadium 2
menjadi stadium 1 , dan stadium 1 menjadi sem ##bu ##h total ? Jika bisa ,
bagaimana cara pen ##go ##batan ##nya ? Mo ##hon pen ##jela ##san ##nya . Ter
##ima kas ##ih [SEP]

[illegible][illegible][illegible]

masuk angin , bu ##ang air dan lain - lain ? Sia ##ng itu me ##mang dia men ##gel ##uhkan Kepala P ##using , Bu ##ang air , mun ##tah , dan dada ses ##ak . Kami kira dia masuk angin , jadi kami kas ##ih tola ##k angin dan lain - lain . Kemudian dia men ##gel ##uhkan , kala ##u tangan ##nya itu dan ja ##ri - ja ##rin ##ya sul ##it dig ##era ##kkan (Dia be ##rp ##iki ##r ken ##a stroke) . Karena masih mua ##l - mua ##l , saya akhirnya pergi untuk amb ##il salon ##pas , dan teman saya men ##jaga ab ##ang saya . La ##lu saya pulang , Ab ##ang saya sudah tidak ada . Kata ##nya sebelum meninggal itu , Dia ke ##jang - ke ##jang luar biasa . Sud ##ah dibawa di rumah sakit , dilakukan CP ##R , namun tidak ter ##sel ##amat ##kan . Beliau me ##mang memiliki ri ##way ##at Darah Ren ##dah . Dari semua ini , apa ##kah ada pen ##jela ##san yang saling berkaitan antara ge ##jala - ge ##jala yang disebut ##kan dia ##tas dok [SEP]

```
INFO:tensorflow:input_ids: 101 67679 85566 117 15595 11889 64981 18049 44725
31585 124 39769 10265 31288 119 52559 12871 51382 10676 76496 85566 10877 10265
71959 33670 10784 38393 10161 11806 23091 119 10685 10237 17674 117 64981 43024
89875 10113 119 89003 118 32105 52559 50274 10486 11806 23091 10265 12729 40796
10116 32500 28977 15290 41585 10676 10659 64951 13908 34675 105676 117 11499
11889 12566 10215 13514 118 13514 136 88468 10376 11910 10911 45306 10671 10588
16039 68637 54087 153 95179 117 11916 11889 12566 117 101833 53538 117 10215
42020 10974 10710 119 87966 32105 10671 34675 105676 117 17760 64985 14399 13187
90470 10174 105676 10215 13514 118 13514 119 52362 10671 10588 16039 68637 117
84844 10138 48371 10676 11910 10215 10201 10401 118 10201 13778 10679 12037
10486 80592 12015 33928 113 18552 10347 33394 20897 10129 67680 10113 57071 114
119 56360 20535 56944 10161 118 56944 10161 117 64981 30448 59159 10782 10559
11030 61658 20084 117 10215 71476 64981 10588 55539 11357 11889 64981 119 10159
11435 64981 107874 117 15595 11889 64981 25147 11868 15290 119 76496 10676 23667
31585 11910 117 18552 11163 37445 118 11163 37445 27120 34384 119 13352 12257
95118 10120 22740 57236 117 28920 40070 11273 117 22736 11868 12718 12912 49158
10706 119 26189 10911 45306 13363 29956 14132 10526 110311 52712 30942 119 35711
23367 10592 117 32500 28977 15290 66558 37142 14434 10265 109002 85783 15345
46503 30216 118 46503 30216 10265 21250 10706 10671 11390 17674 102
```

```
INFO:tensorflow:input_ids: 101 67679 85566 117 15595 11889 64981 18049 44725
31585 124 39769 10265 31288 119 52559 12871 51382 10676 76496 85566 10877 10265
71959 33670 10784 38393 10161 11806 23091 119 10685 10237 17674 117 64981 43024
89875 10113 119 89003 118 32105 52559 50274 10486 11806 23091 10265 12729 40796
10116 32500 28977 15290 41585 10676 10659 64951 13908 34675 105676 117 11499
11889 12566 10215 13514 118 13514 136 88468 10376 11910 10911 45306 10671 10588
16039 68637 54087 153 95179 117 11916 11889 12566 117 101833 53538 117 10215
42020 10974 10710 119 87966 32105 10671 34675 105676 117 17760 64985 14399 13187
90470 10174 105676 10215 13514 118 13514 119 52362 10671 10588 16039 68637 117
84844 10138 48371 10676 11910 10215 10201 10401 118 10201 13778 10679 12037
10486 80592 12015 33928 113 18552 10347 33394 20897 10129 67680 10113 57071 114
119 56360 20535 56944 10161 118 56944 10161 117 64981 30448 59159 10782 10559
11030 61658 20084 117 10215 71476 64981 10588 55539 11357 11889 64981 119 10159
11435 64981 107874 117 15595 11889 64981 25147 11868 15290 119 76496 10676 23667
31585 11910 117 18552 11163 37445 118 11163 37445 27120 34384 119 13352 12257
95118 10120 22740 57236 117 28920 40070 11273 117 22736 11868 12718 12912 49158
```

10706 119 26189 10911 45306 13363 29956 14132 10526 110311 52712 30942 119 35711
23367 10592 117 32500 28977 15290 66558 37142 14434 10265 109002 85783 15345
46503 30216 118 46503 30216 10265 21250 10706 10671 11390 17674 102

[illegible][illegible][illegible][illegible]

INFO:tensorflow:label: gagal-jantung (id = 2)

```
INFO:tensorflow:label: gagal-jantung (id = 2)
```

```
[18]: #Example on first observation in the training set
print("Sentence : ", train_InputExamples.iloc[0].text_a)
print("-"*30)
print("Tokens : ", tokenizer.tokenize(train_InputExamples.iloc[0].text_a))
print("-"*30)
print("Input IDs : ", train_features[0].input_ids)
print("-"*30)
print("Input Masks : ", train_features[0].input_mask)
print("-"*30)
print("Segment IDs : ", train_features[0].segment_ids)
```



```

0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

```

##Creating A Multi-Class Classifier Model

```

[19]: def create_model(is_predicting, input_ids, input_mask, segment_ids, labels,
                        num_labels):

    bert_module = hub.Module(
        BERT_MODEL_HUB,
        trainable=True)
    bert_inputs = dict(
        input_ids=input_ids,
        input_mask=input_mask,
        segment_ids=segment_ids)
    bert_outputs = bert_module(
        inputs=bert_inputs,
        signature="tokens",
        as_dict=True)

    # Use "pooled_output" for classification tasks on an entire sentence.
    # Use "sequence_outputs" for token-level output.
    output_layer = bert_outputs["pooled_output"]

    hidden_size = output_layer.shape[-1].value

    # Create our own layer to tune for politeness data.
    output_weights = tf.get_variable(
        "output_weights", [num_labels, hidden_size],
        initializer=tf.truncated_normal_initializer(stddev=0.02))

    output_bias = tf.get_variable(
        "output_bias", [num_labels], initializer=tf.zeros_initializer())

    with tf.variable_scope("loss"):

        # Dropout helps prevent overfitting
        output_layer = tf.nn.dropout(output_layer, keep_prob=0.9)

        logits = tf.matmul(output_layer, output_weights, transpose_b=True)
        logits = tf.nn.bias_add(logits, output_bias)

```

```

log_probs = tf.nn.log_softmax(logits, axis=-1)

# Convert labels into one-hot encoding
one_hot_labels = tf.one_hot(labels, depth=num_labels, dtype=tf.float32)

predicted_labels = tf.squeeze(tf.argmax(log_probs, axis=-1, output_type=tf.
→int32))
# If we're predicting, we want predicted labels and the probabilities.
if is_predicting:
    return (predicted_labels, log_probs)

# If we're train/eval, compute loss between predicted and actual label
per_example_loss = -tf.reduce_sum(one_hot_labels * log_probs, axis=-1)
loss = tf.reduce_mean(per_example_loss)
return (loss, predicted_labels, log_probs)

```

```

[20]: #A function that adapts our model to work for training, evaluation, and
→prediction.

# model_fn_builder actually creates our model function
# using the passed parameters for num_labels, learning_rate, etc.
def model_fn_builder(num_labels, learning_rate, num_train_steps,
                    num_warmup_steps):
    """Returns `model_fn` closure for TPUEstimator."""
    def model_fn(features, labels, mode, params): # pylint:
→disable=unused-argument
        """The `model_fn` for TPUEstimator."""

        input_ids = features["input_ids"]
        input_mask = features["input_mask"]
        segment_ids = features["segment_ids"]
        label_ids = features["label_ids"]

        is_predicting = (mode == tf.estimator.ModeKeys.PREDICT)

        # TRAIN and EVAL
        if not is_predicting:

            (loss, predicted_labels, log_probs) = create_model(
                is_predicting, input_ids, input_mask, segment_ids, label_ids,
→num_labels)

            train_op = bert.optimization.create_optimizer(
                loss, learning_rate, num_train_steps, num_warmup_steps, use_tpu=False)

            # Calculate evaluation metrics.
            def metric_fn(label_ids, predicted_labels):

```

```

accuracy = tf.metrics.accuracy(label_ids, predicted_labels)
true_pos = tf.metrics.true_positives(
    label_ids,
    predicted_labels)
true_neg = tf.metrics.true_negatives(
    label_ids,
    predicted_labels)
false_pos = tf.metrics.false_positives(
    label_ids,
    predicted_labels)
false_neg = tf.metrics.false_negatives(
    label_ids,
    predicted_labels)

return {
    "eval_accuracy": accuracy,
    "true_positives": true_pos,
    "true_negatives": true_neg,
    "false_positives": false_pos,
    "false_negatives": false_neg
}

eval_metrics = metric_fn(label_ids, predicted_labels)

if mode == tf.estimator.ModeKeys.TRAIN:
    return tf.estimator.EstimatorSpec(mode=mode,
        loss=loss,
        train_op=train_op)
else:
    return tf.estimator.EstimatorSpec(mode=mode,
        loss=loss,
        eval_metric_ops=eval_metrics)
else:
    (predicted_labels, log_probs) = create_model(
        is_predicting, input_ids, input_mask, segment_ids, label_ids,
        num_labels)

    predictions = {
        'probabilities': log_probs,
        'labels': predicted_labels
    }
    return tf.estimator.EstimatorSpec(mode, predictions=predictions)

# Return the actual model function in the closure
return model_fn

```

```
[21]: # Compute train and warmup steps from batch size
# These hyperparameters are copied from this colab notebook (https://colab.
↳sandbox.google.com/github/tensorflow/tpu/blob/master/tools/colab/
↳bert_finetuning_with_cloud_tpus.ipynb)
BATCH_SIZE = 16
LEARNING_RATE = 2e-5
NUM_TRAIN_EPOCHS = 5
# Warmup is a period of time where the learning rate is small and gradually
↳increases--usually helps training.
WARMUP_PROPORTION = 0.1
# Model configs
SAVE_CHECKPOINTS_STEPS = 300
SAVE_SUMMARY_STEPS = 100

# Compute train and warmup steps from batch size
num_train_steps = int(len(train_features) / BATCH_SIZE * NUM_TRAIN_EPOCHS)
num_warmup_steps = int(num_train_steps * WARMUP_PROPORTION)

# Specify output directory and number of checkpoint steps to save
run_config = tf.estimator.RunConfig(
    model_dir=OUTPUT_DIR,
    save_summary_steps=SAVE_SUMMARY_STEPS,
    save_checkpoints_steps=SAVE_CHECKPOINTS_STEPS)

# Specify output directory and number of checkpoint steps to save
run_config = tf.estimator.RunConfig(
    model_dir=OUTPUT_DIR,
    save_summary_steps=SAVE_SUMMARY_STEPS,
    save_checkpoints_steps=SAVE_CHECKPOINTS_STEPS)
```

```
[22]: #Initializing the model and the estimator
```

```
model_fn = model_fn_builder(
    num_labels=len(label_list),
    learning_rate=LEARNING_RATE,
    num_train_steps=num_train_steps,
    num_warmup_steps=num_warmup_steps)

estimator = tf.estimator.Estimator(
    model_fn=model_fn,
    config=run_config,
    params={"batch_size": BATCH_SIZE})
```

```
INFO:tensorflow:Using config: {'_model_dir': '/GD/My Drive/Colab
Notebooks/5epochs', '_tf_random_seed': None, '_save_summary_steps': 100,
'_save_checkpoints_steps': 300, '_save_checkpoints_secs': None,
'_session_config': allow_soft_placement: true
graph_options {
```

```

    rewrite_options {
      meta_optimizer_iterations: ONE
    }
  }
, '_keep_checkpoint_max': 5, '_keep_checkpoint_every_n_hours': 10000,
'_log_step_count_steps': 100, '_train_distribute': None, '_device_fn': None,
'_protocol': None, '_eval_distribute': None, '_experimental_distribute': None,
'_experimental_max_worker_delay_secs': None, '_session_creation_timeout_secs':
7200, '_service': None, '_cluster_spec':
<tensorflow.python.training.server_lib.ClusterSpec object at 0x7fae6af94da0>,
'_task_type': 'worker', '_task_id': 0, '_global_id_in_cluster': 0, '_master':
'', '_evaluation_master': '', '_is_chief': True, '_num_ps_replicas': 0,
'_num_worker_replicas': 1}

INFO:tensorflow:Using config: {'_model_dir': '/GD/My Drive/Colab
Notebooks/5epochs', '_tf_random_seed': None, '_save_summary_steps': 100,
'_save_checkpoints_steps': 300, '_save_checkpoints_secs': None,
'_session_config': allow_soft_placement: true
graph_options {
  rewrite_options {
    meta_optimizer_iterations: ONE
  }
}
, '_keep_checkpoint_max': 5, '_keep_checkpoint_every_n_hours': 10000,
'_log_step_count_steps': 100, '_train_distribute': None, '_device_fn': None,
'_protocol': None, '_eval_distribute': None, '_experimental_distribute': None,
'_experimental_max_worker_delay_secs': None, '_session_creation_timeout_secs':
7200, '_service': None, '_cluster_spec':
<tensorflow.python.training.server_lib.ClusterSpec object at 0x7fae6af94da0>,
'_task_type': 'worker', '_task_id': 0, '_global_id_in_cluster': 0, '_master':
'', '_evaluation_master': '', '_is_chief': True, '_num_ps_replicas': 0,
'_num_worker_replicas': 1}

```

we will now create an input builder function that takes our training feature set (`train_features`) and produces a generator. This is a pretty standard design pattern for working with Tensorflow Estimators.

```

[23]: # Create an input function for training. drop_remainder = True for using TPUs.
train_input_fn = bert.run_classifier.input_fn_builder(
    features=train_features,
    seq_length=MAX_SEQ_LENGTH,
    is_training=True,
    drop_remainder=False)

# Create an input function for validating. drop_remainder = True for using TPUs.
val_input_fn = run_classifier.input_fn_builder(
    features=val_features,
    seq_length=MAX_SEQ_LENGTH,
    is_training=False,

```

```
drop_remainder=False)
```

```
## Training & Evaluating
```

```
[25]: #Training the model
print(f'Beginning Training!')
current_time = datetime.now()
estimator.train(input_fn=train_input_fn, max_steps=num_train_steps)
print("Training took time ", datetime.now() - current_time)
```

Beginning Training!

INFO:tensorflow:Calling model_fn.

INFO:tensorflow:Calling model_fn.

INFO:tensorflow:Saver not created because there are no variables in the graph to restore

INFO:tensorflow:Saver not created because there are no variables in the graph to restore

/usr/local/lib/python3.6/dist-packages/tensorflow_core/python/framework/indexed_slices.py:424: UserWarning: Converting sparse IndexedSlices to a dense Tensor of unknown shape. This may consume a large amount of memory.

"Converting sparse IndexedSlices to a dense Tensor of unknown shape. "

INFO:tensorflow:Done calling model_fn.

INFO:tensorflow:Done calling model_fn.

INFO:tensorflow:Create CheckpointSaverHook.

INFO:tensorflow:Create CheckpointSaverHook.

INFO:tensorflow:Graph was finalized.

INFO:tensorflow:Graph was finalized.

INFO:tensorflow:Restoring parameters from /GD/My Drive/Colab Notebooks/5epochs/model.ckpt-0

INFO:tensorflow:Restoring parameters from /GD/My Drive/Colab Notebooks/5epochs/model.ckpt-0

WARNING:tensorflow:From /usr/local/lib/python3.6/dist-packages/tensorflow_core/python/training/saver.py:1069: get_checkpoint_mtimes (from tensorflow.python.training.checkpoint_management) is deprecated and will be removed in a future version.

Instructions for updating:

Use standard file utilities to get mtimes.

WARNING:tensorflow:From /usr/local/lib/python3.6/dist-packages/tensorflow_core/python/training/saver.py:1069: get_checkpoint_mtimes (from tensorflow.python.training.checkpoint_management) is deprecated and will

be removed in a future version.
Instructions for updating:
Use standard file utilities to get mtimes.

INFO:tensorflow:Running local_init_op.
INFO:tensorflow:Running local_init_op.
INFO:tensorflow:Done running local_init_op.
INFO:tensorflow:Done running local_init_op.

INFO:tensorflow:Saving checkpoints for 0 into /GD/My Drive/Colab Notebooks/5epochs/model.ckpt.
INFO:tensorflow:Saving checkpoints for 0 into /GD/My Drive/Colab Notebooks/5epochs/model.ckpt.

INFO:tensorflow:loss = 1.6099598, step = 0
INFO:tensorflow:loss = 1.6099598, step = 0
INFO:tensorflow:global_step/sec: 0.540929
INFO:tensorflow:global_step/sec: 0.540929
INFO:tensorflow:loss = 1.2701113, step = 100 (184.874 sec)
INFO:tensorflow:loss = 1.2701113, step = 100 (184.874 sec)
INFO:tensorflow:global_step/sec: 0.599952
INFO:tensorflow:global_step/sec: 0.599952
INFO:tensorflow:loss = 0.5506887, step = 200 (166.677 sec)
INFO:tensorflow:loss = 0.5506887, step = 200 (166.677 sec)

INFO:tensorflow:Saving checkpoints for 300 into /GD/My Drive/Colab Notebooks/5epochs/model.ckpt.
INFO:tensorflow:Saving checkpoints for 300 into /GD/My Drive/Colab Notebooks/5epochs/model.ckpt.

INFO:tensorflow:global_step/sec: 0.510864
INFO:tensorflow:global_step/sec: 0.510864
INFO:tensorflow:loss = 0.13547128, step = 300 (195.748 sec)
INFO:tensorflow:loss = 0.13547128, step = 300 (195.748 sec)

INFO:tensorflow:Saving checkpoints for 325 into /GD/My Drive/Colab Notebooks/5epochs/model.ckpt.
INFO:tensorflow:Saving checkpoints for 325 into /GD/My Drive/Colab Notebooks/5epochs/model.ckpt.

INFO:tensorflow:Loss for final step: 0.26931295.
INFO:tensorflow:Loss for final step: 0.26931295.

Training took time 0:11:19.488675

```
[26]: #Evaluating the model with Validation set  
eval_results = estimator.evaluate(input_fn=val_input_fn, steps=None)
```

INFO:tensorflow:Calling model_fn.

INFO:tensorflow:Calling model_fn.

INFO:tensorflow:Saver not created because there are no variables in the graph to restore

INFO:tensorflow:Saver not created because there are no variables in the graph to restore

/usr/local/lib/python3.6/dist-packages/tensorflow_core/python/framework/indexed_slices.py:424: UserWarning: Converting sparse IndexedSlices to a dense Tensor of unknown shape. This may consume a large amount of memory.
"Converting sparse IndexedSlices to a dense Tensor of unknown shape. "

INFO:tensorflow:Done calling model_fn.

INFO:tensorflow:Done calling model_fn.

INFO:tensorflow:Starting evaluation at 2020-07-29T19:39:25Z

INFO:tensorflow:Starting evaluation at 2020-07-29T19:39:25Z

INFO:tensorflow:Graph was finalized.

INFO:tensorflow:Graph was finalized.

INFO:tensorflow:Restoring parameters from /GD/My Drive/Colab Notebooks/5epochs/model.ckpt-325

INFO:tensorflow:Restoring parameters from /GD/My Drive/Colab Notebooks/5epochs/model.ckpt-325

INFO:tensorflow:Running local_init_op.

INFO:tensorflow:Running local_init_op.

INFO:tensorflow:Done running local_init_op.

INFO:tensorflow:Done running local_init_op.

INFO:tensorflow:Finished evaluation at 2020-07-29-19:39:43

INFO:tensorflow:Finished evaluation at 2020-07-29-19:39:43

INFO:tensorflow:Saving dict for global step 325: eval_accuracy = 0.7307692, false_negatives = 12.0, false_positives = 12.0, global_step = 325, loss = 0.689952, true_negatives = 43.0, true_positives = 193.0

INFO:tensorflow:Saving dict for global step 325: eval_accuracy = 0.7307692, false_negatives = 12.0, false_positives = 12.0, global_step = 325, loss = 0.689952, true_negatives = 43.0, true_positives = 193.0

```
INFO:tensorflow:Saving 'checkpoint_path' summary for global step 325: /GD/My Drive/Colab Notebooks/5epochs/model.ckpt-325
```

```
INFO:tensorflow:Saving 'checkpoint_path' summary for global step 325: /GD/My Drive/Colab Notebooks/5epochs/model.ckpt-325
```

```
[27]: eval_results
```

```
[27]: {'eval_accuracy': 0.7307692,  
      'false_negatives': 12.0,  
      'false_positives': 12.0,  
      'global_step': 325,  
      'loss': 0.689952,  
      'true_negatives': 43.0,  
      'true_positives': 193.0}
```

```
[28]: predictions = estimator.predict(val_input_fn)
```

```
[29]: preds_result = []  
      for prediction in predictions:  
          preds_result.append((prediction['probabilities'], prediction['labels']))
```

```
INFO:tensorflow:Calling model_fn.
```

```
INFO:tensorflow:Calling model_fn.
```

```
INFO:tensorflow:Saver not created because there are no variables in the graph to restore
```

```
INFO:tensorflow:Saver not created because there are no variables in the graph to restore
```

```
INFO:tensorflow:Done calling model_fn.
```

```
INFO:tensorflow:Done calling model_fn.
```

```
INFO:tensorflow:Graph was finalized.
```

```
INFO:tensorflow:Graph was finalized.
```

```
INFO:tensorflow:Restoring parameters from /GD/My Drive/Colab Notebooks/5epochs/model.ckpt-325
```

```
INFO:tensorflow:Restoring parameters from /GD/My Drive/Colab Notebooks/5epochs/model.ckpt-325
```

```
INFO:tensorflow:Running local_init_op.
```

```
INFO:tensorflow:Running local_init_op.
```

```
INFO:tensorflow:Done running local_init_op.
```

```
INFO:tensorflow:Done running local_init_op.
```

```
[30]: y_pred = list(map(lambda x: x[1], preds_result))
```

```
[31]: mapping = dict()

for i in range(len(label_list)):
    mapping[label_list[i]] = i

y_actual = list(map(lambda x: mapping[x], val['category'].tolist()))
```

```
[32]: from sklearn.metrics import confusion_matrix

confusion_matrix(y_actual, y_pred)
```

```
[32]: array([[45,  1,  2,  4,  3],
           [ 2, 42,  3,  2, 10],
           [ 4,  3, 29,  2, 11],
           [ 5,  0,  1, 43,  1],
           [ 0,  8,  5,  0, 34]])
```

```
[33]: val_pred = val.copy()
val_pred['pred'] = list(map(lambda x: label_list[x], y_pred))
val_pred.to_csv('prediction_final_raw.csv')
```

```
[34]: from sklearn.metrics import classification_report
print(classification_report(val_pred['category'], val_pred['pred']))
```

	precision	recall	f1-score	support
aritmia	0.78	0.71	0.74	59
gagal-jantung	0.72	0.59	0.65	49
hipertensi	0.80	0.82	0.81	55
serangan-jantung	0.58	0.72	0.64	47
stroke	0.84	0.86	0.85	50
accuracy			0.74	260
macro avg	0.75	0.74	0.74	260
weighted avg	0.75	0.74	0.74	260

```
[35]: val_pred.head()
```

```
[35]:
```

	category	...	pred
1001	serangan-jantung	...	aritmia
1266	stroke	...	stroke
503	gagal-jantung	...	serangan-jantung
756	hipertensi	...	hipertensi
459	gagal-jantung	...	stroke

[5 rows x 7 columns]

#Reference: Most of the code has been taken from the following resource:

- https://colab.research.google.com/github/google-research/bert/blob/master/predicting_movie_reviews_w