Sentiment Analysis-CuDNNLSTM

January 20, 2020

1 Sentiment Analysis with an RNN

Run in Google Colab

View source on GitHub

http://www.polyvista.com/blog/wp-content/uploads/2015/06/sentiment-customer-exp-large.png

1.1 What is Sentiment Analysis?

Sentiment Analysis also know as opinion mining refers to the identification, extraction and study of sentiment states by using natural language processing, text analysis, computational linguistics and biometrics.

1.2 Sentiment Analysis with an Recurrent Neural Network

We will use a RNN for sentiment analysis because we care for the sequence in the data.

1.2.1 Imports

[1]:
```
import re
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt

from tensorflow.keras.models import Sequential, load_model
from tensorflow.compat.v1.keras.layers import CuDNNLSTM, Embedding,␣
 ↪Dropout,Dense
from tensorflow.keras.preprocessing.text import Tokenizer
from tensorflow.keras.preprocessing.sequence import pad_sequences
from tensorflow.keras.optimizers import RMSprop


# import keras
# from keras.models import Sequential, load_model
# from keras.layers import Dense, Embedding, Dropout
# from keras.preprocessing.text import Tokenizer
# from keras.preprocessing.sequence import pad_sequences
import tensorflow as tf
from tensorflow.python.client import device_lib
```

[2]:

```
from tensorflow.compat.v1 import ConfigProto, InteractiveSession



config = ConfigProto()
config.gpu_options.per_process_gpu_memory_fraction = 0.6
config.gpu_options.allow_growth = True
session = InteractiveSession(config=config)
```

[3]:

```
from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "all" #This is for multiple print statements per cell
```

[4]:

```
value = tf.test.is_gpu_available(
    cuda_only=False,
    min_cuda_compute_capability=None
)
print ('***If TF can access GPU: ***\n\n',value) # MUST RETURN True IF IT CAN!!
```

```
WARNING:tensorflow:From <ipython-input-4-cb50da41978a>:3: is_gpu_available (from
tensorflow.python.framework.test_util) is deprecated and will be removed in a
future version.
Instructions for updating:
Use `tf.config.list_physical_devices('GPU')` instead.
***If TF can access GPU: ***

 True
```

[5]:

```
value = tf.config.list_physical_devices('GPU')
print(value)
```

```
[PhysicalDevice(name='/physical_device:GPU:0', device_type='GPU')]
```

[6]:

```
print(device_lib.list_local_devices())
```

```
[name: "/device:CPU:0"
device_type: "CPU"
memory_limit: 268435456
locality {
}
incarnation: 10262187590866292540
, name: "/device:XLA_CPU:0"
device_type: "XLA_CPU"
memory_limit: 17179869184
locality {
}
incarnation: 16202977451625880926
physical_device_desc: "device: XLA_CPU device"
, name: "/device:XLA_GPU:0"
device_type: "XLA_GPU"
memory_limit: 17179869184
locality {
}
incarnation: 3792806479844820577
physical_device_desc: "device: XLA_GPU device"
, name: "/device:GPU:0"
device_type: "GPU"
memory_limit: 1259942707
locality {
  bus_id: 1
  links {
  }
}
incarnation: 6442080958398937588
physical_device_desc: "device: 0, name: GeForce MX150, pci bus id: 0000:02:00.0,
compute capability: 6.1"
]
```

[7]:
```
tf.debugging.set_log_device_placement(True)
```

[8]:
```
 tf
print("Num GPUs Available: ", len(tf.config.experimental.
 ↪list_physical_devices('GPU')))
```

[8]:
```
<module 'tensorflow' from '/home/erolerten/anaconda3/envs/venv-
tensorflow/lib/python3.7/site-packages/tensorflow/__init__.py'>

Num GPUs Available:  1
```

2 Place tensors on the CPU

3 with tf.device(‘/GPU:0’):

a = tf.constant([[1.0, 2.0, 3.0], [4.0, 5.0, 6.0]]) b = tf.constant([[1.0, 2.0], [3.0, 4.0], [5.0, 6.0]])

c = tf.matmul(a, b) print(c)

3.0.1 Loading in Dataset

[9]:

```
data1 = pd.read_csv('Tweets.csv')
data2 = pd.read_csv('stanford-tweets.csv',sep=',')
# data1 = data1.sample(frac=1).reset_index(drop=True)
# data2 = data2.sample(frac=1).reset_index(drop=True)
print(data1.shape)
print(data2.shape)

data1.head()
data2.head()
```

```
(14640, 15)
(1600000, 2)
```

[9]:

```
             tweet_id airline_sentiment  airline_sentiment_confidence  \
0  570306133677760513           neutral                        1.0000
1  570301130888122368          positive                        0.3486
2  570301083672813571           neutral                        0.6837
3  570301031407624196          negative                        1.0000
4  570300817074462722          negative                        1.0000

  negativereason  negativereason_confidence         airline  \
0            NaN                        NaN  Virgin America
1            NaN                     0.0000  Virgin America
2            NaN                        NaN  Virgin America
3     Bad Flight                     0.7033  Virgin America
4     Can't Tell                     1.0000  Virgin America

  airline_sentiment_gold        name negativereason_gold  retweet_count  \
0                    NaN     cairdin                 NaN              0
1                    NaN    jnardino                 NaN              0
2                    NaN  yvonnalynn                 NaN              0
3                    NaN    jnardino                 NaN              0
4                    NaN    jnardino                 NaN              0

                                                text tweet_coord  \
0                @VirginAmerica What @dhepburn said.         NaN
1  @VirginAmerica plus you've added commercials t…         NaN
2  @VirginAmerica I didn't today… Must mean I n…         NaN
3  @VirginAmerica it's really aggressive to blast…         NaN
4  @VirginAmerica and it's a really big bad thing…         NaN

               tweet_created tweet_location               user_timezone
0  2015-02-24 11:35:52 -0800            NaN  Eastern Time (US & Canada)
1  2015-02-24 11:15:59 -0800            NaN  Pacific Time (US & Canada)
2  2015-02-24 11:15:48 -0800      Lets Play  Central Time (US & Canada)
```

```
3  2015-02-24 11:15:36 -0800            NaN  Pacific Time (US & Canada)
4  2015-02-24 11:14:45 -0800            NaN  Pacific Time (US & Canada)
```

```
[9]:   sentiment                                              text
     0  negative  @switchfoot http://twitpic.com/2y1zl - Awww, t…
     1  negative  is upset that he can't update his Facebook by …
     2  negative  @Kenichan I dived many times for the ball. Man…
     3  negative    my whole body feels itchy and like its on fire
     4  negative  @nationwideclass no, it's not behaving at all…
```

Removing all columns except the airline_sentiment and text column.

```
[10]: data1 = data1[['airline_sentiment', 'text']]
      new_columns = ['sentiment','text']
      data1.columns = new_columns
      data1.head()
```

```
[10]:   sentiment                                              text
      0   neutral                 @VirginAmerica What @dhepburn said.
      1  positive  @VirginAmerica plus you've added commercials t…
      2   neutral  @VirginAmerica I didn't today… Must mean I n…
      3  negative  @VirginAmerica it's really aggressive to blast…
      4  negative  @VirginAmerica and it's a really big bad thing…
```

```
[11]: df = data1.append(data2, ignore_index = True)
      print(df.shape)
      df
```

```
(1614640, 2)
```

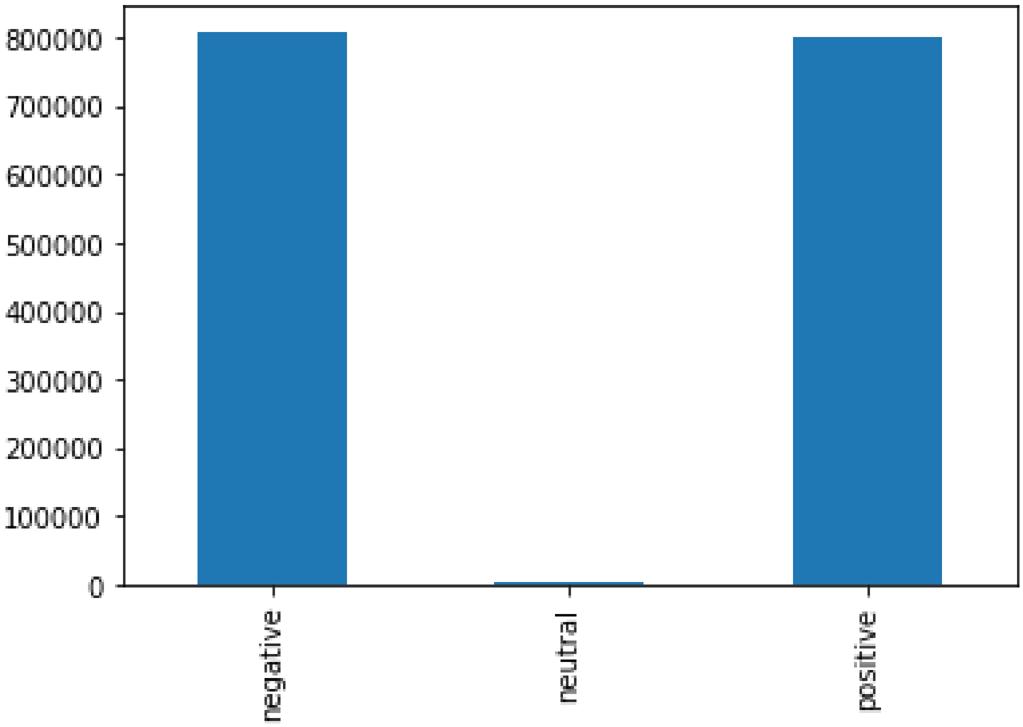
```
[11]:         sentiment                                              text
      0         neutral                 @VirginAmerica What @dhepburn said.
      1        positive  @VirginAmerica plus you've added commercials t…
      2         neutral  @VirginAmerica I didn't today… Must mean I n…
      3        negative  @VirginAmerica it's really aggressive to blast…
      4        negative  @VirginAmerica and it's a really big bad thing…
      …             …                                                 …
      1614635  positive  Just woke up. Having no school is the best fee…
      1614636  positive  TheWDB.com - Very cool to hear old Walt interv…
      1614637  positive  Are you ready for your MoJo Makeover? Ask me f…
      1614638  positive  Happy 38th Birthday to my boo of alll time!!! …
      1614639  positive  happy #charitytuesday @theNSPCC @SparksCharity…

      [1614640 rows x 2 columns]
```

3.0.2 Data exploration

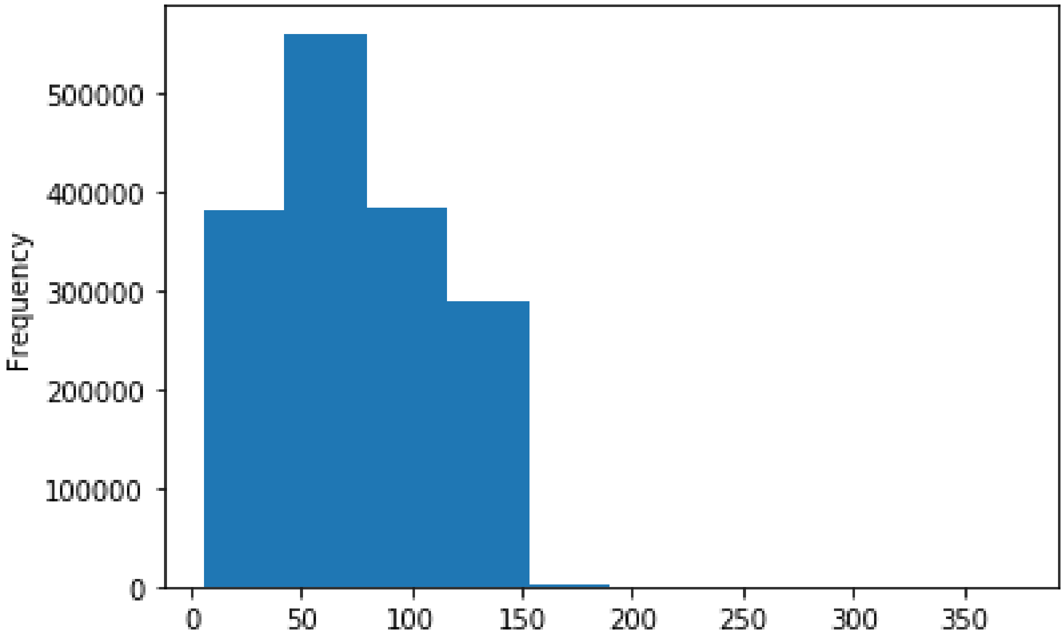
```
[12]: df['sentiment'].value_counts().sort_index().plot.bar()
```

```
[12]: <matplotlib.axes._subplots.AxesSubplot at 0x7fcf40187bd0>
```



```
[13]: df['text'].str.len().plot.hist()
```

```
[13]: <matplotlib.axes._subplots.AxesSubplot at 0x7fcf407b5310>
```



3.0.3 Preprocessing

[14]:
```
# data['text'] = data['text'].str.replace('@VirginAmerica', '')
# data.head()
df = df.sample(frac=1).reset_index(drop=True)
df
```

[14]:
```
        sentiment                                                text
0        negative                   Complaning to paypal  The owe me.
1        positive  MÆ°a, thoáº£i mÃ¡i vÃ£i. Tá» i nay cÃ³ khá» i …
2        negative                  The weather is really upsetting me
3        positive  Gara2 salah jalan, I arrived so late at the Vi…
4        positive  @eight7teen Hey wanna hear how to get the bank…
…           …                                                   …
1614635  positive  @kybabe1001 where r u sitting?? allison and i …
1614636  negative  Hi I'm emma, I love you and you dont know my n…
1614637  positive                                    So far, so good.
1614638  negative  @SassySenna *text* I know  Thats a great idea,…
1614639  positive          just got 80 bucks for disneyland tomorrow!

[1614640 rows x 2 columns]
```

[15]:
```
df['text'].apply(lambda x: x.lower()) #transform text to lowercase
df['text'] = df['text'].apply(lambda x: re.sub('[^a-zA-z0-9\s]', '', x))
```

```
df['text'].head()
```

```
[15]: 0                          complaning to paypal  the owe me.
      1          mæ°a, thoáº£i mã¡i vã£i. tá» i nay cã³ khá» i …
      2                        the weather is really upsetting me
      3          gara2 salah jalan, i arrived so late at the vi…
      4          @eight7teen hey wanna hear how to get the bank…
                                       …
      1614635    @kybabe1001 where r u sitting?? allison and i …
      1614636    hi i'm emma, i love you and you dont know my n…
      1614637                                    so far, so good.
      1614638    @sassysenna *text* i know  thats a great idea,…
      1614639          just got 80 bucks for disneyland tomorrow!
      Name: text, Length: 1614640, dtype: object
```

```
[15]: 0                     Complaning to paypal  The owe me
      1                  Ma thoi mi vi Ti nay c khi th  l m
      2                  The weather is really upsetting me
      3    Gara2 salah jalan I arrived so late at the Vit…
      4    eight7teen Hey wanna hear how to get the bank …
      Name: text, dtype: object
```

```
[16]: df['sentiment']
```

```
[16]: 0          negative
      1          positive
      2          negative
      3          positive
      4          positive
                   …
      1614635    positive
      1614636    negative
      1614637    positive
      1614638    negative
      1614639    positive
      Name: sentiment, Length: 1614640, dtype: object
```

```
[17]: df = df[df['sentiment'] != 'neutral']
```

```
[18]: df
```

```
[18]:         sentiment                                                      text
      0        negative                    Complaning to paypal  The owe me
      1        positive                 Ma thoi mi vi Ti nay c khi th  l m
      2        negative                 The weather is really upsetting me
      3        positive  Gara2 salah jalan I arrived so late at the Vit…
      4        positive  eight7teen Hey wanna hear how to get the bank …
```

```
…           …                                                   …
1614635  positive  kybabe1001 where r u sitting allison and i r g…
1614636  negative  Hi Im emma I love you and you dont know my nam…
1614637  positive                                    So far so good
1614638  negative  SassySenna text I know  Thats a great idea let…
1614639  positive         just got 80 bucks for disneyland tomorrow

[1611541 rows x 2 columns]
```

[19]:
```
# from numba import jit, cuda
```

[20]:
```
tokenizer = Tokenizer(num_words=5000, split=" ")
tokenizer.fit_on_texts(df['text'].values)

X = tokenizer.texts_to_sequences(df['text'].values)
X = pad_sequences(X) # padding our text vector so they all have the same length
X[:5]
```

[20]:
```
array([[   0,    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
           0,    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
           0,    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
           0,    0,    0,    2,    3, 3364,   15],
       [   0,    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
           0,    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
           0,    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
           0, 1234, 1732,  675, 2261,  964,  800],
       [   0,    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
           0,    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
           0,    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
           0,    0,    3,  271,    8,   61,   15],
       [   0,    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
           0,    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
           0,    0,    0,    0,    1, 1323,   16,  296,   23,    3, 4902,
        1267,   28,  459,  165,   33,  212,  926],
       [   0,    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
           0,    0,    0,    0,    0,    0,    0,    0,  152,  166,
         273,   71,    2,   32,    3, 1268,   15,  257,    6,   98,    7,
         257,    6,   92,   86,   56,   22,  548]], dtype=int32)
```

3.0.4 Creating model

[32]:
```
model = Sequential()
model.add(Embedding(5000, 256, input_length=X.shape[1]))
model.add(Dropout(0.3))
model.add(CuDNNLSTM(256, return_sequences=True))
model.add(Dropout(0.3))
model.add(CuDNNLSTM(256))
model.add(Dropout(0.3))
model.add(Dense(2, activation='softmax'))
```

```
Executing op VarHandleOp in device /job:localhost/replica:0/task:0/device:GPU:0
Executing op VarHandleOp in device /job:localhost/replica:0/task:0/device:GPU:0
```

[33]:

```
optimizer_RMS = RMSprop(learning_rate=0.001, rho=0.9)

model.compile(loss='binary_crossentropy', optimizer=optimizer_RMS,␣
 ↪metrics=['accuracy'])
model.summary()
```

```
Model: "sequential_3"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_3 (Embedding)      (None, 40, 256)           1280000
_________________________________________________________________
dropout_11 (Dropout)         (None, 40, 256)           0
_________________________________________________________________
cu_dnnlstm_8 (CuDNNLSTM)     (None, 40, 256)           526336
_________________________________________________________________
dropout_12 (Dropout)         (None, 40, 256)           0
_________________________________________________________________
cu_dnnlstm_9 (CuDNNLSTM)     (None, 256)               526336
_________________________________________________________________
dropout_13 (Dropout)         (None, 256)               0
_________________________________________________________________
dense_3 (Dense)              (None, 2)                 514
=================================================================
Total params: 2,333,186
Trainable params: 2,333,186
Non-trainable params: 0
_________________________________________________________________
```

[34]:

```
y = pd.get_dummies(df['sentiment']).values
[print(df['sentiment'][i], y[i]) for i in range(0,5)]
```

```
negative [1 0]
positive [0 1]
negative [1 0]
positive [0 1]
positive [0 1]
```

[34]:

```
[None, None, None, None, None]
```

[35]:

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,␣
 ↪random_state=0)
```

3.0.5 Training model

[36]:

```
batch_size = 32
epochs = 8

import time

start = time.time()
model.fit(X_train, y_train, epochs=epochs, batch_size=batch_size, verbose=2)
end = time.time()
elapsed = end - start
print(elapsed/60," minutes")
```

```
Executing op RangeDataset in device /job:localhost/replica:0/task:0/device:CPU:0
Executing op RepeatDataset in device
/job:localhost/replica:0/task:0/device:CPU:0
Executing op MapDataset in device /job:localhost/replica:0/task:0/device:CPU:0
Executing op PrefetchDataset in device
/job:localhost/replica:0/task:0/device:CPU:0
Executing op FlatMapDataset in device
/job:localhost/replica:0/task:0/device:CPU:0
Executing op TensorDataset in device
/job:localhost/replica:0/task:0/device:CPU:0
Executing op RepeatDataset in device
/job:localhost/replica:0/task:0/device:CPU:0
Executing op ZipDataset in device /job:localhost/replica:0/task:0/device:CPU:0
Executing op ParallelMapDataset in device
/job:localhost/replica:0/task:0/device:CPU:0
Executing op DatasetCardinality in device
/job:localhost/replica:0/task:0/device:CPU:0
Train on 1289232 samples
Epoch 1/8
Executing op ModelDataset in device /job:localhost/replica:0/task:0/device:CPU:0
Executing op AnonymousIteratorV2 in device
/job:localhost/replica:0/task:0/device:CPU:0
Executing op MakeIterator in device /job:localhost/replica:0/task:0/device:CPU:0
Executing op VarHandleOp in device /job:localhost/replica:0/task:0/device:GPU:0
Executing op AssignVariableOp in device
/job:localhost/replica:0/task:0/device:GPU:0
Executing op VarHandleOp in device /job:localhost/replica:0/task:0/device:GPU:0
Executing op VarHandleOp in device /job:localhost/replica:0/task:0/device:GPU:0
Executing op VarHandleOp in device /job:localhost/replica:0/task:0/device:GPU:0
Executing op VarHandleOp in device /job:localhost/replica:0/task:0/device:GPU:0
Executing op VarHandleOp in device /job:localhost/replica:0/task:0/device:CPU:0
Executing op VarHandleOp in device /job:localhost/replica:0/task:0/device:GPU:0
Executing op LogicalNot in device /job:localhost/replica:0/task:0/device:GPU:0
Executing op Assert in device /job:localhost/replica:0/task:0/device:GPU:0
Executing op VarHandleOp in device /job:localhost/replica:0/task:0/device:GPU:0
Executing op VarHandleOp in device /job:localhost/replica:0/task:0/device:GPU:0
Executing op VarHandleOp in device /job:localhost/replica:0/task:0/device:GPU:0
Executing op VarHandleOp in device /job:localhost/replica:0/task:0/device:GPU:0
Executing op VarHandleOp in device /job:localhost/replica:0/task:0/device:GPU:0
Executing op VarHandleOp in device /job:localhost/replica:0/task:0/device:GPU:0
Executing op VarHandleOp in device /job:localhost/replica:0/task:0/device:GPU:0
Executing op __inference_distributed_function_4730 in device
/job:localhost/replica:0/task:0/device:GPU:0
1289232/1289232 - 1157s - loss: 0.4570 - accuracy: 0.7909
Epoch 2/8
1289232/1289232 - 1149s - loss: 0.4460 - accuracy: 0.8000
Epoch 3/8
1289232/1289232 - 1148s - loss: 0.4441 - accuracy: 0.8015
Epoch 4/8
1289232/1289232 - 1147s - loss: 0.4443 - accuracy: 0.8017
Epoch 5/8
1289232/1289232 - 1149s - loss: 0.4439 - accuracy: 0.8019
Epoch 6/8
1289232/1289232 - 1153s - loss: 0.4451 - accuracy: 0.8021
Epoch 7/8
1289232/1289232 - 1156s - loss: 0.4456 - accuracy: 0.8019
Epoch 8/8
1289232/1289232 - 1156s - loss: 0.4453 - accuracy: 0.8017
Executing op DeleteIterator in device
/job:localhost/replica:0/task:0/device:CPU:0
```

```
[36]: <tensorflow.python.keras.callbacks.History at 0x7fce30a4c590>
```

```
153.57674520413082  minutes
```

```
[37]: model.save('sentiment_analysis-20012020.h5')
```

```
Executing op ReadVariableOp in device
/job:localhost/replica:0/task:0/device:GPU:0
Executing op Identity in device /job:localhost/replica:0/task:0/device:GPU:0
Executing op ReadVariableOp in device
/job:localhost/replica:0/task:0/device:CPU:0
Executing op ReadVariableOp in device
/job:localhost/replica:0/task:0/device:GPU:0
Executing op Identity in device /job:localhost/replica:0/task:0/device:GPU:0
```

3.0.6 Testing model

[38]:
```
predictions = model.predict(X_test)

[print(df['text'][i], predictions[i], y_test[i]) for i in range(0, 5)]
```

```
Executing op RangeDataset in device /job:localhost/replica:0/task:0/device:CPU:0
Executing op RepeatDataset in device
/job:localhost/replica:0/task:0/device:CPU:0
Executing op MapDataset in device /job:localhost/replica:0/task:0/device:CPU:0
Executing op PrefetchDataset in device
/job:localhost/replica:0/task:0/device:CPU:0
Executing op FlatMapDataset in device
/job:localhost/replica:0/task:0/device:CPU:0
Executing op TensorDataset in device
/job:localhost/replica:0/task:0/device:CPU:0
Executing op RepeatDataset in device
/job:localhost/replica:0/task:0/device:CPU:0
Executing op ZipDataset in device /job:localhost/replica:0/task:0/device:CPU:0
Executing op ParallelMapDataset in device
/job:localhost/replica:0/task:0/device:CPU:0
Executing op ModelDataset in device /job:localhost/replica:0/task:0/device:CPU:0
Executing op AnonymousIteratorV2 in device
/job:localhost/replica:0/task:0/device:CPU:0
Executing op __inference_distributed_function_972077 in device
/job:localhost/replica:0/task:0/device:GPU:0
Complaning to paypal  The owe me [0.39116168 0.60883826]  [0 1]
Ma thoi mi vi Ti nay c khi th  l m  [0.5083627  0.49163726] [1 0]
The weather is really upsetting me  [0.6681618  0.33183816] [1 0]
Gara2 salah jalan I arrived so late at the Vita Charm event Now sitting among
FDers who all look positively gorgeous  Koukla [0.53400517 0.4659948 ]  [1 0]
eight7teen Hey wanna hear how to get the bank acct Wire me 1  and ill wire you 1
and then they will be open  49694 [0.91013384 0.08986621]  [1 0]
```

[38]:
```
[None, None, None, None, None]
```

[39]:
```
accurate_prediction_count, inaccurate_prediction_count = 0, 0
for i, prediction in enumerate(predictions):
    if np.argmax(prediction)==np.argmax(y_test[i]):
        accurate_prediction_count += 1
    else:
        inaccurate_prediction_count += 1

total_predictions = accurate_prediction_count + inaccurate_prediction_count
print('Number of prediprinttns: ', total_predictions)
print('Number of accurate predictions: ', accurate_prediction_count)
print('Number of false predictions: ', inaccurate_prediction_count)
```

```python
print('Accuracy: ', accurate_prediction_count/total_predictions)
```

```
Number of prediprinttns:  322309
Number of accurate predictions:  259218
Number of false predictions:  63091
Accuracy:  0.8042530615030917
```

[40]:

```python
pos_count, neu_count, neg_count = 0, 0, 0
real_pos, real_neu, real_neg = 0, 0, 0
for i, prediction in enumerate(predictions):
    if np.argmax(prediction)==2:
        pos_count += 1
    elif np.argmax(prediction)==1:
        neu_count += 1
    else:
        neg_count += 1

    if np.argmax(y_test[i])==2:
        real_pos += 1
    elif np.argmax(y_test[i])==1:
        real_neu += 1
    else:
        real_neg +=1

print('Positive predictions:', pos_count)
print('Neutral predictions:', neu_count)
print('Negative predictions:', neg_count)
print('Real positive:', real_pos)
print('Real neutral:', real_neu)
print('Real negative:', real_neg)
```

```
Positive predictions: 0
Neutral predictions: 151088
Negative predictions: 171221
Real positive: 0
Real neutral: 160271
Real negative: 162038
```

3.1 Improvements we could implement

Weight classes (because data is skew)

Train more epochs

Use bigger network

Try other word number

3.2 Resources

Recurrent Neural Networks Explained (my own post and video)

Sentiment Analysis (Wikipedia)

What is the best way to do sentiment analysis with Python? (Quora)

How to Do Sentiment Analysis (Siraj Raval)