Sentiment_Analysis-binary-classification-BRNN-LSTM

January 24, 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

[2]:
```
import re
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt

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

[3]:
```
from tensorflow.compat.v1 import ConfigProto
from tensorflow.compat.v1 import InteractiveSession
```

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

[4]:
```
from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "all" #This is for multiple print␣
↪statements per cell
```

[5]:
```
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-5-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
```

[6]:
```
value = tf.config.list_physical_devices('GPU')
print(value)
```

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

[7]:
```
print(device_lib.list_local_devices())
```

```
[name: "/device:CPU:0"
device_type: "CPU"
memory_limit: 268435456
locality {
}
incarnation: 8971790019553799407
, name: "/device:XLA_CPU:0"
device_type: "XLA_CPU"
memory_limit: 17179869184
locality {
}
incarnation: 15506684212406741406
physical_device_desc: "device: XLA_CPU device"
, name: "/device:XLA_GPU:0"
device_type: "XLA_GPU"
memory_limit: 17179869184
```

```
locality {
}
incarnation: 2744856512113414661
physical_device_desc: "device: XLA_GPU device"
, name: "/device:GPU:0"
device_type: "GPU"
memory_limit: 1259942707
locality {
  bus_id: 1
  links {
  }
}
incarnation: 10681043894120664247
physical_device_desc: "device: 0, name: GeForce MX150, pci bus id: 0000:02:00.0,
compute capability: 6.1"
]
```

[8]:
```
tf.debugging.set_log_device_placement(True)
```

[9]:
```
 tf
print("Num GPUs Available: ", len(tf.config.experimental.
 ↪list_physical_devices('GPU')))
```

[9]:
```
<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

[10]:
```
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)
```

[10]:

	tweet_id	airline_sentiment	airline_sentiment_confidence	negativereason	negativereason_confidence	airline	airline_sentiment_gold	name	negativereason_gold	retweet_count	text	tweet_coord	tweet_created	tweet_location	user_timezone
0	570306133677760513	neutral	1.0000	NaN	NaN	Virgin America	NaN	cairdin	NaN	0	@VirginAmerica What @dhepburn said.	NaN	2015-02-24 11:35:52 -0800	NaN	Eastern Time (US & Canada)
1	570301130888122368	positive	0.3486	NaN	0.0000	Virgin America	NaN	jnardino	NaN	0	@VirginAmerica plus you've added commercials t…	NaN	2015-02-24 11:15:59 -0800	NaN	Pacific Time (US & Canada)
2	570301083672813571	neutral	0.6837	NaN	NaN	Virgin America	NaN	yvonnalynn	NaN	0	@VirginAmerica I didn't today… Must mean I n…	NaN	2015-02-24 11:15:48 -0800	Lets Play	Central Time (US & Canada)
3	570301031407624196	negative	1.0000	Bad Flight	0.7033	Virgin America	NaN	jnardino	NaN	0	@VirginAmerica it's really aggressive to blast…	NaN	2015-02-24 11:15:36 -0800	NaN	Pacific Time (US & Canada)
4	570300817074462722	negative	1.0000	Can't Tell	1.0000	Virgin America	NaN	jnardino	NaN	0	@VirginAmerica and it's a really big bad thing…	NaN	2015-02-24 11:14:45 -0800	NaN	Pacific Time (US & Canada)

[10]:

	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.

[11]:
```
data1 = data1[['airline_sentiment', 'text']]
new_columns = ['sentiment','text']
data1.columns = new_columns
data1.head()
```

[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…

[12]:
```
df = data1.append(data2, ignore_index = True)
print(df.shape)
df
```

```
(1614640, 2)
```

[12]:

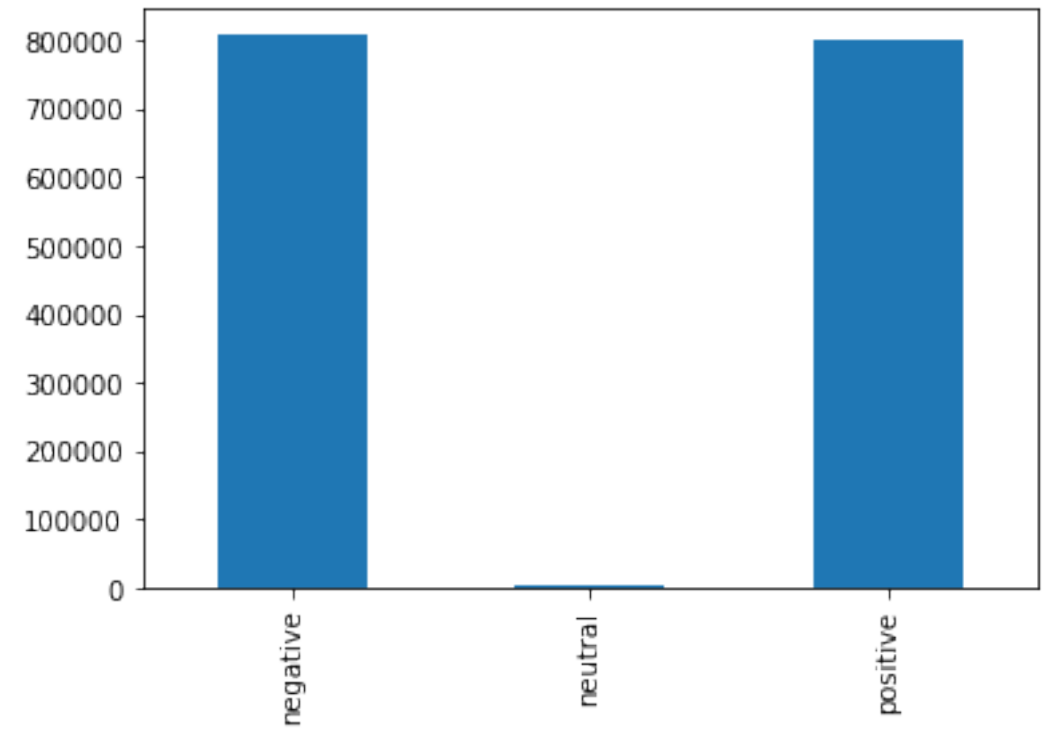
	sentiment	text
0	neutral	@VirginAmerica What @dhepburn said.
1	positive	@VirginAmerica plus you've added commercials t…
2	neutral	@VirginAmerica I didn't today… Must mean I n…
3	negative	@VirginAmerica it's really aggressive to blast…
4	negative	@VirginAmerica and it's a really big bad thing…
…	…	…
1614635	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

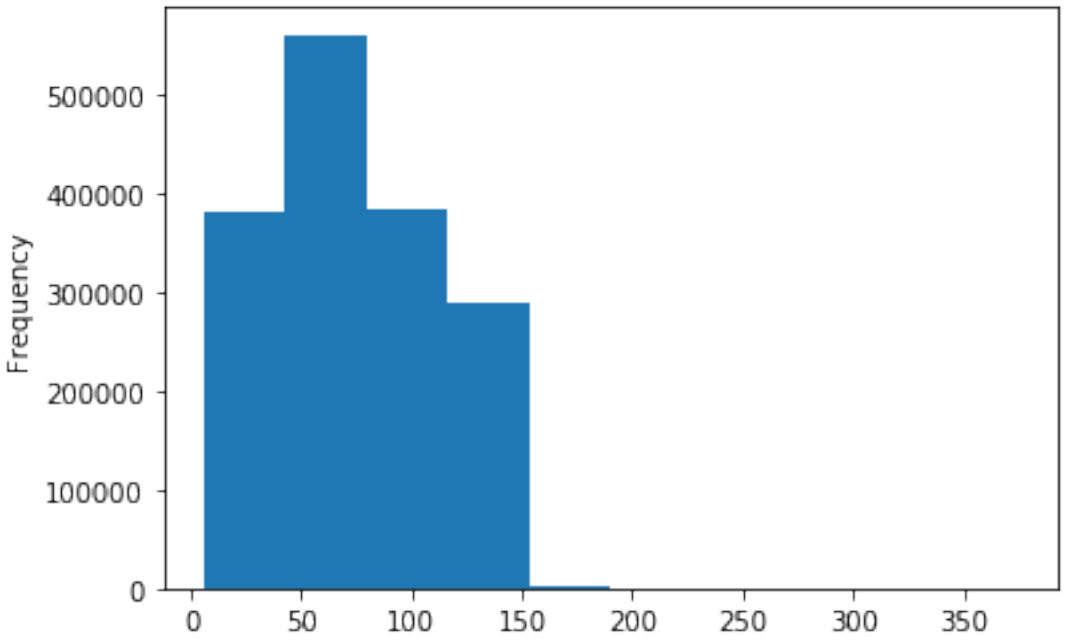
[13]:
```
df['sentiment'].value_counts().sort_index().plot.bar()
```

[13]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f09fb86b210>
```



[14]:
```
df['text'].str.len().plot.hist()
```

[14]: <matplotlib.axes._subplots.AxesSubplot at 0x7f09fc75d4d0>



3.0.3 Preprocessing

[15]:
```
# data['text'] = data['text'].str.replace('@VirginAmerica', '')
# data.head()
df = df.sample(frac=1).reset_index(drop=True)
df
```

[15]:
```
         sentiment                                                text
0         negative   @noeljohan that same place my $2 is  nonexistent
1         positive                                          going home
2         negative  @scigirl543 I wasn't, lol my parents are, I sa…
3         positive  @femme_ecarlate You guys must have done that F…
4         negative  i dread having the dreams i've always wanted t…
…            …                                                   …
1614635   positive  @ShermanHu the last thing a woman needs is ano…
1614636   negative  @KingFOE not one single person  lol these peop…
1614637   positive  @Jessicaveronica Jess! Australia's such a babe…
1614638   negative  u wouldn't think it'd be this hard to find sho…
1614639   negative  My little girl is 7 weeks tomorrow… that's n…

[1614640 rows x 2 columns]
```

[16]:
```
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()
```

[16]:
```
0           @noeljohan that same place my $2 is  nonexistent
1                                                going home
2          @scigirl543 i wasn't, lol my parents are, i sa…
3          @femme_ecarlate you guys must have done that f…
4          i dread having the dreams i've always wanted t…
                               …
1614635    @shermanhu the last thing a woman needs is ano…
1614636    @kingfoe not one single person  lol these peop…
1614637    @jessicaveronica jess! australia's such a babe…
1614638    u wouldn't think it'd be this hard to find sho…
1614639    my little girl is 7 weeks tomorrow… that's n…
Name: text, Length: 1614640, dtype: object
```

[16]:
```
0       noeljohan that same place my 2 is  nonexistent
1                                          going home
2    scigirl543 I wasnt lol my parents are I saw a …
3    femme_ecarlate You guys must have done that Fr…
4    i dread having the dreams ive always wanted to…
```

```
Name: text, dtype: object
```

[17]:
```
df['sentiment']
```

[17]:
```
0          negative
1          positive
2          negative
3          positive
4          negative
             …
1614635    positive
1614636    negative
1614637    positive
1614638    negative
1614639    negative
Name: sentiment, Length: 1614640, dtype: object
```

[18]:
```
df = df[df['sentiment'] != 'neutral']
```

[19]:
```
df
```

[19]:
```
         sentiment                                                text
0         negative     noeljohan that same place my 2 is  nonexistent
1         positive                                          going home
2         negative  scigirl543 I wasnt lol my parents are I saw a …
3         positive  femme_ecarlate You guys must have done that Fr…
4         negative  i dread having the dreams ive always wanted to…
…            …                                                    …
1614635   positive  ShermanHu the last thing a woman needs is anot…
1614636   negative  KingFOE not one single person  lol these peopl…
1614637   positive  Jessicaveronica Jess Australias such a babe co…
1614638   negative  u wouldnt think itd be this hard to find showe…
1614639   negative  My little girl is 7 weeks tomorrow thats nearl…

[1611541 rows x 2 columns]
```

[20]:
```
# from numba import jit, cuda
```

[21]:
```
vocabulary_size = 5000

tokenizer = Tokenizer(num_words=vocabulary_size, 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]
```

[21]:
```
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,   18,  244,  408,    5,   80,    8],
       [   0,    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
           0,    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
           0,    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
           0,    0,    0,    0,    0,   44,   76],
       [   0,    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
           0,    0,    0,    0,    0,    0,    1,  388,   52,    5,  793,
          35,    1,  279,    4,  245,   20,   11, 2646,    9,   26,   20,
           4,  238,    7,   58,   13,   25,   18],
       [   0,    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
           0,    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
           7,  198,  324,   17,  186,   18,  326,   78,    1,  388,   79,
         335,    1,  136,   58,   10,    3,  665],
       [   0,    0,    0,    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,  173,    3,  747,  132,  190,
         355,    2,   17,  385,    3, 1578,    4]], dtype=int32)
```

3.0.4 Creating model

[22]:
```
model = Sequential()
model.add(Embedding(vocabulary_size, 256, input_length=X.shape[1]))
model.add(Dropout(0.3))
model.add(Bidirectional(LSTM(256, return_sequences=True, dropout=0.3,␣
 ↪recurrent_dropout=0.2)))
model.add(Bidirectional(LSTM(256, dropout=0.3, recurrent_dropout=0.2)))
model.add(Dense(2, activation='sigmoid'))
```

```
Executing op RandomUniform in device
/job:localhost/replica:0/task:0/device:GPU:0
Executing op Sub in device /job:localhost/replica:0/task:0/device:GPU:0
Executing op Mul in device /job:localhost/replica:0/task:0/device:GPU:0
Executing op Add 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 VarIsInitializedOp 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 AssignVariableOp in device
/job:localhost/replica:0/task:0/device:GPU:0
Executing op RandomUniform 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 Fill in device /job:localhost/replica:0/task:0/device:GPU:0
Executing op ConcatV2 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 RandomUniform in device
/job:localhost/replica:0/task:0/device:GPU:0
Executing op RandomUniform 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 RandomUniform in device
/job:localhost/replica:0/task:0/device:GPU:0
Executing op RandomUniform 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
```

[23]:
```
model.compile(loss='binary_crossentropy', optimizer='adam',␣
 ↪metrics=['accuracy'])
model.summary()
```

```
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
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, 40, 256)           1280000
_________________________________________________________________
dropout_1 (Dropout)          (None, 40, 256)           0
_________________________________________________________________
bidirectional_1 (Bidirection (None, 40, 512)           1050624
_________________________________________________________________
bidirectional_2 (Bidirection (None, 512)               1574912
_________________________________________________________________
dense_1 (Dense)              (None, 2)                 1026
=================================================================
Total params: 3,906,562
Trainable params: 3,906,562
Non-trainable params: 0
_________________________________________________________________
```

[24]:
```
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]
negative [1 0]
```

[24]:
```
[None, None, None, None, None]
```

[25]:
```
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

[26]:
```
batch_size = 32
epochs = 7

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 Reshape 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

/home/erolerten/anaconda3/envs/venv-tensorflow/lib/python3.7/site-
packages/tensorflow_core/python/framework/indexed_slices.py:433: 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. "

Epoch 1/7
Executing op __inference_keras_scratch_graph_6441 in device
/job:localhost/replica:0/task:0/device:GPU:0
 - 5162s - loss: 0.4251 - accuracy: 0.8029
Epoch 2/7
 - 5148s - loss: 0.3968 - accuracy: 0.8196
Epoch 3/7
 - 5150s - loss: 0.3900 - accuracy: 0.8232
Epoch 4/7
 - 5139s - loss: 0.3878 - accuracy: 0.8245
Epoch 5/7
 - 5140s - loss: 0.3872 - accuracy: 0.8247
Epoch 6/7
 - 5139s - loss: 0.3876 - accuracy: 0.8246
Epoch 7/7
 - 5138s - loss: 0.3884 - accuracy: 0.8242
```

[26]: <keras.callbacks.callbacks.History at 0x7f08e01a2250>

```
600.2875208417574  minutes
```

[27]:
```
model.save('sentiment_analysis-23012020.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:GPU:0
Executing op Identity in device /job:localhost/replica:0/task:0/device:GPU:0
```

3.0.6 Testing model

[28]:
```
predictions = model.predict(X_test)

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

```
Executing op __inference_keras_scratch_graph_1699226 in device
/job:localhost/replica:0/task:0/device:GPU:0
noeljohan that same place my 2 is  nonexistent [0.00956778 0.9904322 ] [0 1]
going home  [0.03006512 0.9699349 ] [0 1]
scigirl543 I wasnt lol my parents are I saw a bit just in passing it was just a
guess You know Im not that cultured  [0.00804161 0.9919585 ] [0 1]
femme_ecarlate You guys must have done that Friday when I wasnt there  Anything
I should know for the final [0.01810214 0.9818981 ] [0 1]
i dread having the dreams ive always wanted to have cause the frigging holidays
a frigging barrier  [0.93713975 0.06286056] [0 1]
```

[28]: [None, None, None, None, None]

[29]:
```
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 predictions: ', total_predictions)
print('Number of accurate predictions: ', accurate_prediction_count)
print('Number of false predictions: ', inaccurate_prediction_count)
print('Accuracy: ', accurate_prediction_count/total_predictions)
```

```
Number of predictions:  322309
Number of accurate predictions:  265676
```

```
Number of false predictions:  56633
Accuracy:  0.8242897343853259
```

[ ]:
```
# 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)
```

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)