Name: Vaarshinni Reddy Andru
Enroll no: 70042952

Use Case Description:

1. Sentiment Analysis on the Twitter dataset

Programming elements:

1. Basics of LSTM

2. Types of RNN

3. Use case: Sentiment Analysis on the Twitter data set

In class programming:

1. Save the model and use the saved model to predict on new text data (ex, “A lot of good things are happening. We are respected again throughout the world, and that's a great thing.@realDonaldTrump”)

```
[1] import pandas as pd #Basic packages for creating dataframes and loading dataset
    import numpy as np

    import matplotlib.pyplot as plt #Package for visualization

    import re #importing package for Regular expression operations

    from sklearn.model_selection import train_test_split #Package for splitting the data

    from sklearn.preprocessing import LabelEncoder #Package for conversion of categorical to Numerical

    from keras.preprocessing.text import Tokenizer #Tokenization
    from tensorflow.keras.preprocessing.sequence import pad_sequences #Add zeros or crop based on the length
    from keras.models import Sequential #Sequential Neural Network
    from keras.layers import Dense, Embedding, LSTM, SpatialDropout1D #For layers in Neural Network
    from keras.utils.np_utils import to_categorical
```

```
[2] from google.colab import drive
    drive.mount('/content/gdrive')
```

```
Mounted at /content/gdrive
```

```
[5] import pandas as pd

    # Load the dataset as a Pandas DataFrame
    dataset = pd.read_csv('/content/gdrive/My Drive/Sentiment.csv')

    # Select only the necessary columns 'text' and 'sentiment'
    mask = dataset.columns.isin(['text', 'sentiment'])
    data = dataset.loc[:, mask]

    # Keeping only the necessary columns
    data['text'] = data['text'].apply(lambda x: x.lower())
    data['text'] = data['text'].apply((lambda x: re.sub('[^a-zA-z0-9\s]', '', x)))
```

```
<ipython-input-5-d0e745dc69e5>:11: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

```python
for idx, row in data.iterrows():
    row[0] = row[0].replace('rt', ' ') #Removing Retweets
    max_fatures = 2000
tokenizer = Tokenizer(num_words=max_fatures, split=' ') #Maximum words is 2000 to tokenize sentence
tokenizer.fit_on_texts(data['text'].values)
X = tokenizer.texts_to_sequences(data['text'].values) #taking values to feature matrix
X = pad_sequences(X) #Padding the feature matrix

embed_dim = 128 #Dimension of the Embedded layer
lstm_out = 196 #Long short-term memory (LSTM) layer neurons
def createmodel():
    model = Sequential() #Sequential Neural Network
    model.add(Embedding(max_fatures, embed_dim,input_length = X.shape[1])) #input dimension 2000 Neurons, output dimension 128 Neurons
    model.add(LSTM(lstm_out, dropout=0.2, recurrent_dropout=0.2)) #Drop out 20%, 196 output Neurons, recurrent dropout 20%
    model.add(Dense(3,activation='softmax')) #3 output neurons[positive, Neutral, Negative], softmax as activation
    model.compile(loss = 'categorical_crossentropy', optimizer='adam',metrics = ['accuracy']) #Compiling the model
    return model
# print(model.summary())
labelencoder = LabelEncoder() #Applying label Encoding on the label matrix
integer_encoded = labelencoder.fit_transform(data['sentiment']) #fitting the model
y = to_categorical(integer_encoded)
X_train, X_test, Y_train, Y_test = train_test_split(X,y, test_size = 0.33, random_state = 42) #67% training data, 33% test data split
batch_size = 32 #Batch size 32
model = createmodel() #Function call to Sequential Neural Network
model.fit(X_train, Y_train, epochs = 1, batch_size=batch_size, verbose = 2) #verbose the higher, the more messages
score,acc = model.evaluate(X_test,Y_test,verbose=2,batch_size=batch_size) #evaluating the model
print(score)
print(acc)
```

```
291/291 - 35s - loss: 0.8267 - accuracy: 0.6426 - 35s/epoch - 119ms/step
144/144 - 2s - loss: 0.7524 - accuracy: 0.6798 - 2s/epoch - 14ms/step
0.7523683905601501
0.6797728538513184
```

[7]

```python
print(model.metrics_names) #metrics of the model
```

```
['loss', 'accuracy']
```

[9]

```python
#1. Save the model and use the saved model to predict on new text data (ex, "A lot of good things are happening. We are respected again throughout the world, and that's a great thing.@realDonaldTrump")
model.save('sentimentAnalysis.h5') #Saving the model
```

[10]

```python
from keras.models import load_model #Importing the package for importing the saved model
model= load_model('sentimentAnalysis.h5') #loading the saved model
```

```python
print(integer_encoded)
print(data['sentiment'])
```

```
[1 2 1 ... 2 0 2]
0         Neutral
1        Positive
2         Neutral
3        Positive
4        Positive
           ...
13866    Negative
13867    Positive
13868    Positive
13869    Negative
13870    Positive
Name: sentiment, Length: 13871, dtype: object
```

```python
# Predicting on the text data
sentence = ['A lot of good things are happening. We are respected again throughout the world, and that is a great thing.@realDonal
sentence = tokenizer.texts_to_sequences(sentence) # Tokenizing the sentence
sentence = pad_sequences(sentence, maxlen=28, dtype='int32', value=0) # Padding the sentence
sentiment_probs = model.predict(sentence, batch_size=1, verbose=2)[0] # Predicting the sentence text
sentiment = np.argmax(sentiment_probs)

print(sentiment_probs)
if sentiment == 0:
    print("Neutral")
elif sentiment < 0:
    print("Negative")
elif sentiment > 0:
    print("Positive")
else:
    print("Cannot be determined")
```

```
1/1 - 0s - 212ms/epoch - 212ms/step
[0.5888069  0.15338774 0.25780532]
Neutral
```

2. Apply GridSearchCV on the source code provided in the class

```
[13] #2. Apply GridSearchCV on the source code provided in the class
```

```
from keras.wrappers.scikit_learn import KerasClassifier #importing Keras classifier
from sklearn.model_selection import GridSearchCV #importing Grid search CV

model = KerasClassifier(build_fn=createmodel,verbose=2) #initiating model to test performance by applying multiple hyper parameters
batch_size= [10, 20, 40] #hyper parameter batch_size
epochs = [1, 2] #hyper parameter no. of epochs
param_grid= {'batch_size':batch_size, 'epochs':epochs} #creating dictionary for batch size, no. of epochs
grid  = GridSearchCV(estimator=model, param_grid=param_grid) #Applying dictionary with hyper parameters
grid_result= grid.fit(X_train,Y_train) #Fitting the model
# summarize results
print("Best: %f using %s" % (grid_result.best_score_, grid_result.best_params_)) #best score, best hyper parameters
```

```
Epoch 2/2
372/372 - 33s - loss: 0.6859 - accuracy: 0.7069 - 33s/epoch - 88ms/step
Epoch 1/2
186/186 - 23s - loss: 0.8411 - accuracy: 0.6352 - 23s/epoch - 123ms/step
Epoch 2/2
186/186 - 21s - loss: 0.6911 - accuracy: 0.7014 - 21s/epoch - 114ms/step
47/47 - 1s - loss: 0.7380 - accuracy: 0.6756 - 929ms/epoch - 20ms/step
Epoch 1/2
186/186 - 23s - loss: 0.8440 - accuracy: 0.6330 - 23s/epoch - 124ms/step
Epoch 2/2
186/186 - 21s - loss: 0.6821 - accuracy: 0.7096 - 21s/epoch - 114ms/step
47/47 - 1s - loss: 0.8045 - accuracy: 0.6590 - 925ms/epoch - 20ms/step
Epoch 1/2
186/186 - 23s - loss: 0.8482 - accuracy: 0.6328 - 23s/epoch - 126ms/step
Epoch 2/2
186/186 - 20s - loss: 0.6973 - accuracy: 0.7013 - 20s/epoch - 108ms/step
47/47 - 1s - loss: 0.7469 - accuracy: 0.6814 - 919ms/epoch - 20ms/step
Epoch 1/2
186/186 - 24s - loss: 0.8462 - accuracy: 0.6336 - 24s/epoch - 126ms/step
Epoch 2/2
186/186 - 20s - loss: 0.6788 - accuracy: 0.7116 - 20s/epoch - 108ms/step
47/47 - 1s - loss: 0.7888 - accuracy: 0.6615 - 1s/epoch - 22ms/step
Epoch 1/2
930/930 - 80s - loss: 0.8162 - accuracy: 0.6480 - 80s/epoch - 86ms/step
Epoch 2/2
930/930 - 77s - loss: 0.6762 - accuracy: 0.7137 - 77s/epoch - 83ms/step
Best: 0.683309 using {'batch_size': 10, 'epochs': 2}
```