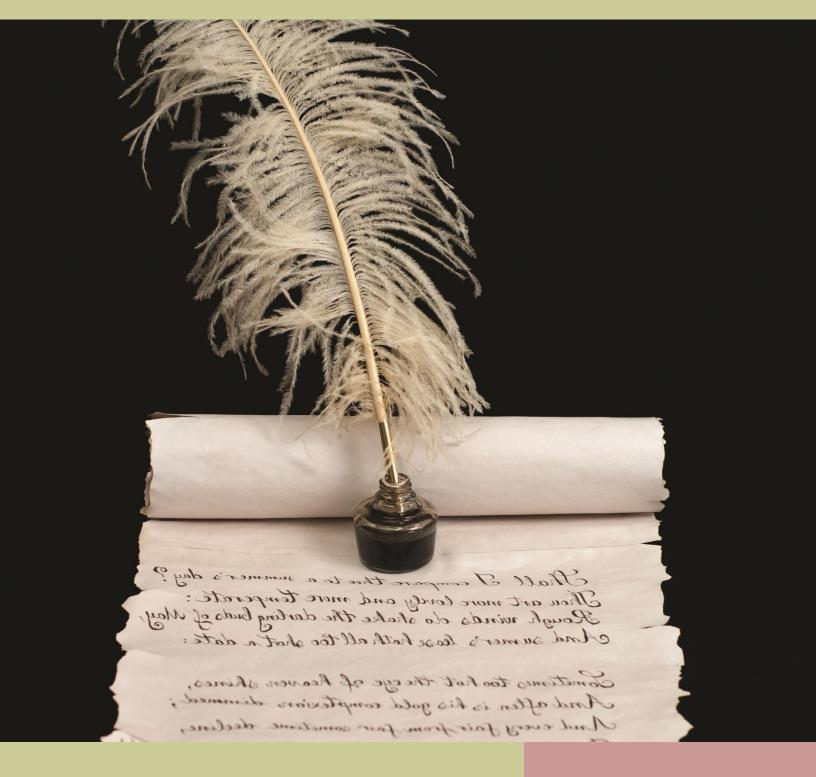
Poem Generation



Neural Network for Poem Generation

Students:
Salem Allosh
Aws Abdul Hamed

Introduction

Main aim of this project is to create a model to generate a poem depending on specific poet pattern.

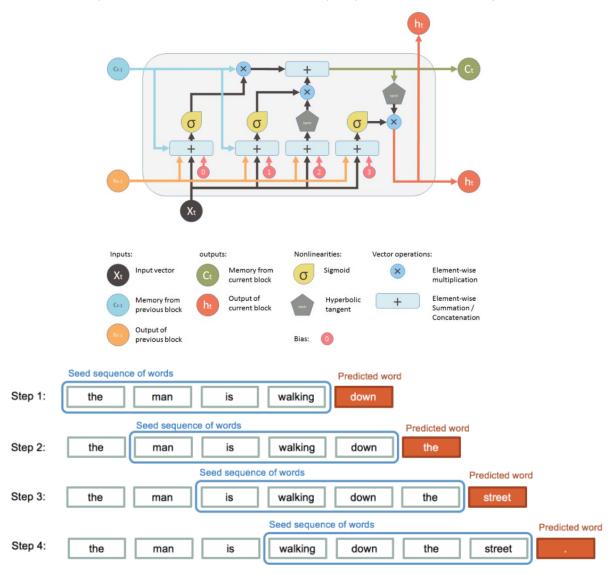
To develop this project, we are going through several steps:

- 1- Choosing the algorithms that achieve the goal.
- 2- Get the Dataset.
- 3- Preprocessing Dataset
- 4- Create network model
- 5- Evaluate the result

Algorithms

Since the sentences in text can be representing in embedding and how the semantics in text over long stretch might be learned using LSTM ..

In text generation, we try to predict the next character or word of the sequence. The text data generally considered as sequence of data. For predicting data in sequence we used deep learning models like RNN or LSTM. LSTM are preferred over RNN in this because of RNN vanishing and exploding gradients problem. Since in text generation we have to memorize large amount of previous data. So for this purpose LSTM are preferred.



Data set

We depended on Arabic_potery data set ,we found it on Kaggle Data set properties:

- Consist of 3 columns:
 - Category :
 - This dataset contain (28) different types of poem ages and types such as:

```
[ ] categories = GetPoemCategories (df) print(categories)

'الإمارات' 'البحرين' 'البخزائر' 'السعودية' 'السودان' 'العراق' 'المغرب']

'اليمن' 'تونس' 'سوريا' 'عمان' 'فلسطين' 'لبنان' 'ليبيا' 'مصر' 'الأردن'

'الكويت' 'قطر' 'موريتانيا' 'العصر الجاهلي' 'العمر الإسلامي'

'العمر العباسي' 'العصر الايوبي' 'العصر الغثماني' 'عصر المخضرمون'

['العصر الاموي' 'العصر الأندنسي' 'العصر المملوكي'
```

- Poet_name:

And contains of 538 poet name

```
🕑 poets = GetPoetNames(df)
  |print(poets)|
   'خلفان بن مصبح' 'سالم بن عبدلله الكراني' 'حماد بن سعيد']
'مبارك بن حمد العقيلي' 'علي بن محمد الشحي' 'محمد بن حمود الشحي'
    'محمد بن صالح المنتفقي' 'يوسف بن رويسم' 'هدى السعدي' 'مانع سعيد العتيبة'
    'عارف الخاجة' 'حمد بن خليفة أبو شهاب' 'كريم معتوق'
    'سالم أبو جمهور القبيسيّ' 'قاسم حداد' 'أحلام مستغانمَي' 'سهام آلُ ٰبراهمَي'
'جلواح' 'عبد القادر الجزائري' 'سليمان بن سحمان' 'عبد الرحمن بن مساعد'
'أحمد اللهيب' 'عبدالرحمن العشماوي' 'شريف بقنه' 'ابن بشير الإحسائي'
    'سامي المالكي' 'غازي القصيبي' 'محمد بن عبود العمودي'
    'مدثر بن إبراهيم بن الحجاز' 'محمد عبد الباري' 'إدريس جمّاع'
    'حمزة الملك طُمبل' 'احمّد مطر' 'نازك الملائكة' 'بدر شاكر السياب'
    'محمد مهدي الجواهري' 'عبد الرزاق عبد الواحد' 'بهاء الدين الصيادي'
    'معروف الرصافي' 'أبو الفيف الكتاني' 'مَحمد اسموني' 'شاعر الحمراء'
    'عبدلله البردوني' 'عبد الولي الشميري' 'ابن شهاب العلوي' 'محمد الشوكاني'
      'ابن طاهر' 'ابن رشيق القيرواني' 'الشاذلي خزنه دار' 'أُبو القاسم الُشابي
    ر و القادر أنزار قباني' 'خليل مردم بك' 'بطرس كرامة'
    'أبو الهدى الصيادي' 'ابن شيخان السالمي' 'سيّف الرحبي' 'صالح الفّهدي'
    'هلال بن سعيد العماني' 'توفيق زياد' 'جورج جريس َفرح' 'محمود دروْيش'
'بديع القشاعلة' 'تميم البرغوثي' 'أسامه محمد زامل' 'إبراهيم طوقان'
```

- Poem_text:
 - Consists of 54767 poem text

```
] SyrianPoem = SplitByCategory(df,"سهوركا")
 SyrianPoem.head()
          category poet name
                                                                            poem text
  4061
                سليم عبدالقادر سوريا
                                                   ...كلُّ وجهٍ منهم⊓\إنهم يطلعون من كلِّ أَفْق
                 سليم عبدالقادر سوريا
                                                          ...فِيn\ الْ لِيْ: مِرْ، وَرَاحَ يَعْدُو أَمَامِي
  4062
                 ... والموت يرقص لي ١٨ماض ، وأعرف ما دربي وما هدفي سليم عبدالقادر سوريا
   4063
                          سليم عبدالقادر
                                                            ...قَلَىn\ لِمَاذَا تَخَافُونَنِي، لَسُتُ أَدُرِي
  4064
                 سوريا
                سليم عبدالقادر سوريا
                                                           ... هَذَا الْحِصِنَارُ الْرُّ هَبِثُ الْمُتَعِبُ الْقَاسِي
   4065
```

Preprocessing

since the data set is large and there is not a sutiable hardware to train such model , so we will split our model to parts and train a model for generate a poem dpending on poet name , to achive that we developed many functions to help us in preprocessing:

Load data set from google drive

```
from pydrive.auth import GoogleAuth
from pydrive.drive import GoogleDrive
from google.colab import auth
from oauth2client.client import GoogleCredentials

[3] auth.authenticate user()
gauth = GoogleAuth()
gauth.credentials = GoogleCredentials.get_application_default()
drive = GoogleDrive(gauth)

[5] fileDownloaded = drive.CreateFile({'id':'1_vfqCzvNaR_ApgaVwCCqbwDa21s95W5T'})
[6] fileDownloaded.GetContentFile('Arabic_poetry_dataset.csv')
```

Load Dataset To Data Frame

- With ensure that the data set has no empty cell

```
import numpy as np
import pandas as pd

missing_values = ["missing",np.nan]
df = pd.read_csv('Arabic_poetry_dataset.csv',na_values= missing_values,usecols = ['poet_name','poem_text','category'])
```

- Head Example



GetPoetNames(DataFrame)

 Take One arguments DataFrame And return all poets names in this dataframe

```
def GetPoetNames(DataFrame):
   poet_name = DataFrame.poet_name.unique()
   return poet_name
```

- GetPoemCategory(DataFrame)
 - Take One arguments DataFrame And return all Categories names in this data frame

```
def GetPoemCategories(Dataframe):
   poem_categories = Dataframe.category.unique()
   return poem_categories
```

GetPoetsNamesByCategories(DataFrame, Category)

 Take Two arguments Dataframe and Category and returen all poets that belong to Category

```
def GetPoetsNamesByCategories(Dataframe, Category):
   PoemCategory = Dataframe.loc[Dataframe.category==Category]
   poetName = PoemCategory.poet_name.unique()
   return poetName
```

- Example:

```
("سوريا", "GetPoetsNamesByCategories(df,")

array(['نطرس كرامة'], 'خليل مردم بك', 'بطرس كرامة'], dtype=object)
```

- GetCategoryByPoetName(DataFrame,PoetName)
 - Take Two arguments Dataframe and Category and returen all
 Category that contain poems for that Poet

```
def GetCategoryByPoetName(Dataframe,PoetName):
   PoetName = Dataframe.loc[Dataframe.poet_name==PoetName]
   CategoryName = PoetName.category.unique()
   return CategoryName
```

- Example:

```
GetCategoryByPoetName(df,"إبِلْيا ابو ماضي")
array(['لبنان'], dtype=object)
```

- SplitByCategory(DataFrame, Category)
 - Create a new Dataset that only contains poems and poets of that belong to Category

```
def SplitByCategory(Dataframe, CategoryName):
    SubDataframe = Dataframe.loc[Dataframe["category"]==CategoryName]
    SubDataframe.to_csv(CategoryName+'.csv',encoding='utf-8-sig')
    return SubDataframe
```

- Example :

```
SyrianPoem = SplitByCategory(df, "سوريا")
SyrianPoem.head()
          category poet_name
                                                                              poem_text
                                                   ...كلُّ وجِـهِ منهمN|إنهم يطلعون من كلُّ أَفْقَ
 4061
                سليم عبدالقادر سوريا
                                                          ...فِي ١٦ ال لِئي: سِرْ، وَرَاحَ يَعْدُو أَمَامِي
                سليم عبدالقادر سوريا
 4062
                ... والموت يرقص لي ١٨ماض ، وأعرف ما دربي وما هدفي سليم عبدالقادر سوريا
 4063
                                                            ...قُلَّى ١٨ لِمَاذَا تَخَافُونَنِي، لَسُتُ أَنْرِي
                سليم عبدالقادر سوريا
 4064
                سليم عبدالقادر سوريا
                                                           ... هَذَا الْحِصْنَالُ الرَّهِيثِ الْمُثْعِثِ الْقَاسِي
 4065
```

SplitByPoetName(DataFrame, PoetName)

 Create a new Dataset that only contains data that belong to PoetName.

```
def SplitByPoetName(Dataframe , PoetName ):
    SubDataframe = Dataframe.loc[Dataframe["poet_name"]==PoetName]
    SubDataframe.to_csv(PoetName+'.csv',encoding='utf-8-sig')
    return SubDataframe
```

- Example:

```
("أبو القاسم السّابي", AboAlqasemDataframe = SplitByPoetName(df
AboAlqasemDataframe.head()
         category poet_name
                                                            poem_text
        ...كأنُّ أاهازِخرُجْنَ من قُرُجاتِ النُّقع داميةً     أبو القاسم الشابي توبس
 3970
               ...بدموع عبدِكَ١٨ما كلُّما بل ربُّما عبثَ البُّكا ۚ أبو القاسم الشابي توبس
 3971
            ... كالنَّمسُر Nمنَاعيشُ رَعْمَ الدَّاءِ والأعداءِ       أبو القاسم الشابي     تونس
 3972
             ...وهُمُومي و١٨ألُّهِما الحُدُّبُّ أنتَ سِرُّ بَلائِي ۚ أبو القاسم الشابي تونس
 3973
               ...نَفْس مِنْ الإِلا أَنْتَ شِعْرى هِلْ لِلَيْلِ الن أبو القاسم الشابي توبس
 3974
EleaAboMadiDataFrame = SplitByPoetName(df, "إبليا ابو ماضي")
EleaAboMadiDataFrame.nunique()
category
                   1
poet_name
                 1
poem text
                172
```

GetPoems(Dataframe)

dtype: int64

 Take One Argument and return a string of all poems in that data frame

```
def GetPoems(Dataframe):
   PoemDF = Dataframe.poem_text
   Poems = ""
   for i in range(len(PoemDF)):
      Poems += PoemDF.iloc[i]
   return Poems
```

```
AboAlqasemPoems = GetPoems(AboAlqasemDataframe)
print(AboAlgasemPoems)
فقدْ فَتُّ في زَنْدِ الدِّيَانَةِ معشّرُ
أتاروا على الْإسلامِ مَنْ قَدْ يُهاجِمُ
فواالحقُّ مَا هذي الزُّوايا وأهلُها
سِوَى مصنع فيهِ تُصاغُ السُّخائِمُ
لَّحَى اللَّهُ مَنْ لَمُ تُسْتَثِّرِه حَمَيُّةُ
على بينه إنْ داهمتَهُ العَظائِمُترجو السُّعادَةَ يِا قَلبي ولو وُجِنتُ
في الكون لم يسَّنعلُ خُزُنٌ ولا أَلْمُ
ولا استحالتُ حياةُ النَّاسِ أجمعُها
وزُلزلتُ هاتِهِ الأَكوانُ والنَّظمُ
فما السُّعادة في الدُّنيا سِوَى جُلُمْ
ناءٍ تُضنَكِّي له أَيَّامَها الأُمَمُّ
ناجتُ به الْأَاسَ أَوِهَامٌ مُعَرْبِدِهُ
لَمَّا تُعَتَّنَّتُهُمُ الأحلامُ والظَّلْمُ
فَهَبُّ كُلُّ لِتَاسِهِ وِيَنْشُدُهُ
 كأنَّما النَّاسُ مَا ناموا ولا حَلْمُوا
```

TokenizePoems(Poem, Print=False)

 Take two arguments String of all poems And a print Boolean Value to print details if want.

- Return:

Corpus: Set of each line in poems

Tokenizer: Tokenizer Object

total_words: number of unrepeated words

len(corpus): number of lines in poems

```
def TokenizePoems(Poem,Print=False):
   tokenizer = Tokenizer()
   corpus = Poem.split("\n")
   tokenizer.fit_on_texts(corpus)
   total_words = len(tokenizer.word_index) + 1
   if(Print):
      print("Words numbers ",total_words)
      print("Tokenizer word index ",tokenizer.word_index)

return corpus,tokenizer,total_words,len(corpus)
```

- Example:

```
Words numbers 11148
Tokenizer word index { '11147 : 'ولْبُلْبِه' ' , 11146 : 'لِخَرُجْنَ مِن فُرُجابَ النَّتِع داميةً ' ] 'ولْبُلْبِها ' : مَا فَى قُرِبُ اللَّهُ عَلَيْه اللَّهُ اللَّهُ اللَّهُ اللَّهُ عَلَيْه اللَّهُ عَلَيْه اللَّهُ عَلَيْه اللَّهُ عَلَيْه اللَّهُ عَلَيْه مِن فُرُجابَ النَّتِع داميةً ' ] ) 'ولاذا السَمالُ مَع الحسَيَّ تسمَّتُ ' والا السَمالُ مَع الحسَيَّ تسمَّتُ ' والا السَمالُ مَع الحسَيِّ السَّمَّةِ والاعداء ' والأعداء ' واللَّهُ السَّمَّ اللَّهُ والاَعداء ' والأَمطار والأَمطار والأَمواء ' والأَمطار والأَمواء ' والمَعلَّم والأَمواء ' والمَعلَّم والأَمطار والأَمواء ' والمَعلَّم والأَمواء ' والمَعلَّم والمَعلَم والمِعلَم والمَعلَم والمُعلَم والمَعلَم والمَعلِ
```

• GenerateInputSequence(Poem, Print)

- Take two arguments String of all poems And a print Boolean Value to print details.
- How it works:

Call TokenizePoems(Poem,Print=False) and receive Corpus Tokanize, TotalWords, len(corpus)

For each corpus assign int value to each word in that corpus using tokenizer.texts_to_sequences([line "From Corpus"])

After getting list of integers that represent the corpus Going to split that list in sequence of consecutive integer list

Retrun

Corpus_num: number of lines line in poems

Tokenizer: Tokenizer Object

total_words : number of unrepeated words
input_sequences: list of consecutive integer

```
def GenerateInputSequence(Poem,Print=False):
    corpus,tokenizer,total_words,curpus_num = TokenizePoems(Poem,Print)
    input_sequences = []
    for line in corpus:
        " مصنوفة من تدالي الكلمات بصبغة النجر #
        token_list = tokenizer.texts_to_sequences([line])[0]
        for i in range(1, len(token_list)):
            n_gram_sequence = token_list[:i+1]
            input_sequences.append(n_gram_sequence)

if(Print):
    print("Corpus length ",corpus)
    print("Corpus Sequence ",input_sequences)
    print("Corpus Sequence Length ",len(input_sequences))

return input_sequences,tokenizer,total_words,curpus_num
```

Eplanation:

Line:	Input Sequences:
[4 2 66 8 67 68 69 70]	[4 2] [4 2 66] [4 2 66 8] [4 2 66 8 67] [4 2 66 8 67 68] [4 2 66 8 67 68 69] [4 2 66 8 67 68 69 70]

Generate_X_Y(Poem, Print=False)

- Take two arguments String of all poems And a print Boolean Value to print details.
- How it works:

Call GenerateInputSequence(Poem,Print) and receive input sequencs And Tokenizer, Total_words And Corpus_num.

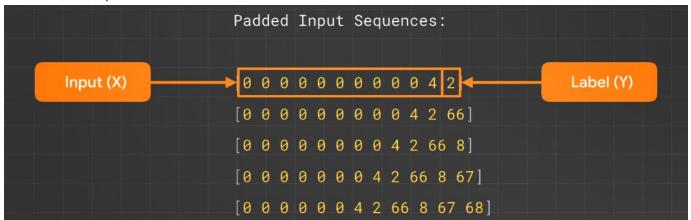
X same As input Sequence But Missing the last item in each sublist. Y represent the last item in each SubList .

Then padding the X to Longest line in Corpus.

Then One Hot Encoding For Y.

```
def Generate_X_Y(Poem,Print=False):
    sequences,tokenizer,words,corpus = GenerateInputSequence(Poem,Print)
    max_sequence_len = max([len(x) for x in sequences])
    sequences = np.array(pad_sequences(sequences, maxlen=max_sequence_len, padding='pre'))
    xs, labels = sequences[:,:-1],sequences[:,-1]
    ys = to_categorical(labels, num_classes=words)
    if(Print):
        print("Max Sequence Length " ,max_sequence_len )
    return xs,ys,tokenizer,words,max_sequence_len
```

- Eplanation :



Build Model

as known there is no rules to build a sufficient NN So we try a several architectures with several Poems.

First try:

Build a neural network consists of 3 Layers:

- 1. Embedding Layer
- 2. Bidirectional(LSTM) layer
- 3. Dense Layer

Hipper Parameters are set as shownSecond try:

```
def BuildModel(PrintDetails=False):
   model = Sequential()
   model.add(Embedding(WordsNumber, 100, input_length=MaxSequenceLength-1))
   model.add(Bidirectional(LSTM(150)))
   model.add(Dense(WordsNumber, activation='softmax'))
   adam = Adam(learning_rate=0.01)
   model.compile(loss='categorical_crossentropy', optimizer=adam, metrics=['accuracy'])
   #earlystop = EarlyStopping(monitor='val_loss', min_delta=0, patience=5, verbose=0, mode='auto')
   if(PrintDetails):|
        print(model.summary())
        tf.keras.utils.plot_model(model, to_file='model.png', show_shapes=True, show_layer_names=True)
   return model
```

Second Try:

Build a neural network consists of 5 Layers:

- 1. Embedding Layer
- 2. LSTM
- 3. Bidirectional(LSTM) layer
- 4. Dense Layer
- 5. Dense Layer

Hipper Parameters are set as shown

```
def BuildModel(PrintDetails=False):
    model = Sequential()
    model.add(Embedding(WordsNumber, 50, input_length=MaxSequenceLength-1))
    model.add(LSTM(100,return_sequences=True))
    model.add(Bidirectional(LSTM(100)))
    model.add(Dense(WordsNumber/2, activation='relu'))
    model.add(Dense(WordsNumber, activation='softmax'))
    adam = Adam(learning_rate=0.01)
    model.compile(loss='categorical_crossentropy', optimizer=adam, metrics=['accuracy'])
    #earlystop = EarlyStopping(monitor='val_loss', min_delta=0, patience=5, verbose=0, mode='auto')
    if(PrintDetails):
        print(model.summary())
        tf.keras.utils.plot_model(model, to_file='model.png', show_shapes=True, show_layer_names=True)
    return model
```

Train And Validation

Elea Abo Madi:

1. First Model

```
Model: "sequential"
D.
   Layer (type)
                                 Output Shape
                                                             Param #
   embedding (Embedding)
                                  (None, 11, 100)
                                                             2232700
   bidirectional (Bidirectional (None, 1000)
                                                             2404000
   dense (Dense)
                                  (None, 22327)
                                                             22349327
   Total params: 26,986,027
   Trainable params: 26,986,027
   Non-trainable params: 0
```

Train with 10 epoch

```
TrainModel2(X,Y,Model2)
Epoch 1/10
1117/1117 [
                          =======] - 114s 99ms/step - loss: 9.6896 - accuracy: 0.0358
Epoch 2/10
1117/1117 [
                                ==] - 109s 98ms/step - loss: 8.6484 - accuracy: 0.0449
Epoch 3/10
1117/1117 [
                                 ==] - 110s 99ms/step - loss: 6.6881 - accuracy: 0.0717
Epoch 4/10
                                ==] - 112s 100ms/step - loss: 4.3872 - accuracy: 0.2393
1117/1117 [
Epoch 5/10
                          =======] - 111s 99ms/step - loss: 2.8664 - accuracy: 0.4406
1117/1117 [
Epoch 6/10
1117/1117 [
                   Epoch 7/10
1117/1117 [
                       ========] - 111s 99ms/step - loss: 1.6484 - accuracy: 0.6467
Epoch 8/10
1117/1117 [
                           =======] - 111s 100ms/step - loss: 1.3998 - accuracy: 0.6943
Epoch 9/10
1117/1117 [
                   =========] - 111s 99ms/step - loss: 1.2673 - accuracy: 0.7173
Epoch 10/10
1117/1117 [==:
```

Accuracy: 0.73

Train time: 2 h

PredectNextOf(tokenizer,Model2,MaxSequenceLength,Sentence = "بيا نفس هذا منزل الاحباب", PredectionLength=20, "بيا نفس هذا منزل الاحباب" = ver/local/lib/python3.7/dist-packages/tensorflow/python/keras/engine/sequential.py:455: UserWarning: warnings.warn('`model.predict_classes()` is deprecated and '

بيا نفس هذا منزل الاحباب أُختَ دار الخُلدِ يا أُمَّ الفُرى الأسى الكُرى وَأَطارَني العاري عَلْيكُمْ ما الكُرى ما فَعْلوا الوَرى تَأَمُّلُ الوَرى صَيَّيًا فَيهِ

PredectNextOf(tokenizer,Model2,MaxSequenceLength,Sentence = "الوقال",PredectionLength=40)

ib/python3.7/dist-packages/tensorflow/python/keras/engine/sequential.py:455: UserWarning: `model.predict_classes()` is deprecated and will be arn('`model.predict_classes()` is deprecated and '

لو قال مِن تُرِيَّ مَرْقُهَا بِيْدِ اللَّرِى مُشَقِّلًا وَأَنْ يَمْتَى وَأَنْ تُسْفَى فَأَمْكِ مِسْتَرَا وَبَعُدُ الوَرِى كُمْ الوَرِى كُمُلْ يَوِهُبُ الوَرِى كَمُلْ الْمَالِيَّ فِي الْحَبِيِّ لِمِينًا لَمُورِى تَعْلَلُ

PredectNextOf(tokenizer,Model2,MaxSequenceLength,Sentence "بصرعن ذا اللب" PredectionLength=40), "predectionLength=40, "predectionLength=40," is deprecated," is deprecated and '(" model.predict_classes() is deprecated and '(" model.predict_classes() is deprecated and '(" model.predict_classes() الفضا الد إثما صورتها هنذ كالمبرّ الذمن فجه الورى لا الذمن الذمن فجه الورى لا الذمن الذمن الذمن الذمن الذمن المناس الذمن الذمن المناس الذمن الدمن المناس الدائمة المناس الذمن المناس الذمن الدائمة المناس الدائمة الدائمة المناس الدائمة الدائمة الدائمة الدائمة المناس الدائمة الدائمة الدائمة المناس الدائمة المناس الدائمة الدائ

2. Second Model

```
Model: "sequential_1"
Layer (type)
                         Output Shape
                                                Param #
______
embedding_1 (Embedding)
                         (None, 11, 50)
                                                 1116350
1stm 1 (LSTM)
                         (None, 11, 100)
                                                 60400
bidirectional_1 (Bidirection (None, 200)
                                                 160800
dense 1 (Dense)
                         (None, 11163)
                                                 2243763
dense 2 (Dense)
                         (None, 22327)
                                                 249258628
Total params: 252,839,941
Trainable params: 252,839,941
Non-trainable params: 0
None
```

```
TrainModel(X,Y,Model)
Epoch 1/10
Epoch 2/10
Epoch 3/10
Epoch 4/10
Epoch 5/10
Epoch 6/10
Epoch 7/10
Epoch 8/10
Epoch 9/10
790/1117 [=========>: ] - ETA: 12:05 - loss: 7.3531 - accuracy: 0.0485
uation
        r,Model,MaxSequenceLength,Sentence = "",PredectionLength=0):
in another tab.
    Show diff
               4h 53m 33s completed at 2:10 PM
```

Google Colab Crushed after 5 hours in waiting So the results will be delivered tomorrow at the session (x)

At the training we can see tat the accuracy is improving So by increasing the epoch the improvement will be grater

Accuracy:

Train time:

To Complete Text: ""