Name: Brij Popatbhai Patel
Student ID: 1146509

Project 2: MI using Word Embedding

Project Description:

In this project, I have implemented classification models that classify the SMS, Emails, or Tweets as spam or ham. The classification task is necessary nowadays because of unimportant and advertised emails and SMS. For this task, I have implemented machine learning as well as deep learning models and to evaluate the models we have used the UCI dataset.

Dataset:

UCI contains a large number of datasets that can be helped to evaluate our projects. For this task, I have used Spambase Data Set provided by the professor. It contains 4601 rows of data with 58 columns that represent features and labels. The last column of the dataset is used to classify data into spam and ham because it has label 1 and 0 that represents spam and not spam(ham) respectively.

Methodology:

1. Use Traditional Machine Learning Classifiers from SKLearn:

To evaluate models, I have used machine learning classifiers and generated the result. For that I have used the following Sklearn libraries:

```
from sklearn.linear_model import LogisticRegression

from sklearn.svm import SVC

from sklearn.naive_bayes import MultinomialNB

from sklearn.tree import DecisionTreeClassifier

from sklearn.neighbors import KNeighborsClassifier

from sklearn.ensemble import RandomForestClassifier
```

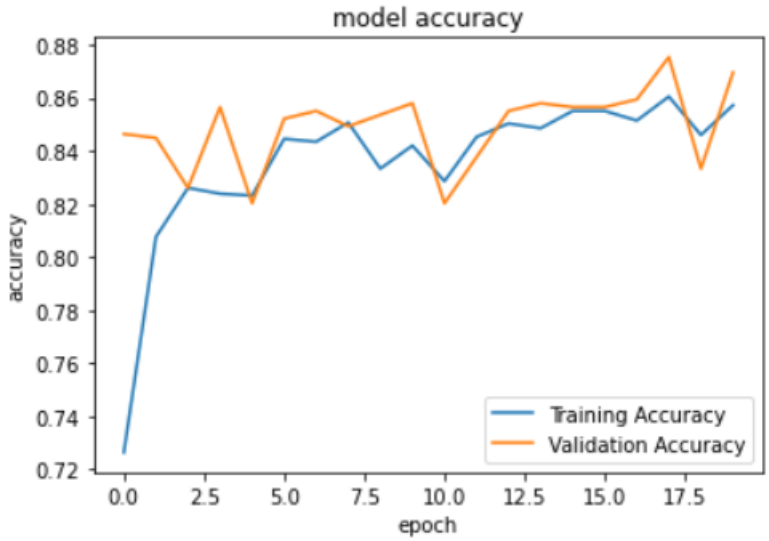
Firstly, I generate the training and testing data and fit the training data into the classifier, and predict testing data. And I have calculated the accuracy score for each classification technique and the result of it are as below:

```
Accuracy Score of Machine Learning Models:
----------------------------------------------------------------------
Logistic Regression:        87.66 %
Support Vector Classifier:  92.7 %
Multinomial Naive Bayes:    88.1 %
Decision Tree:              90.27 %
KNeighbors Classifier:      89.4 %
Random Forest Classifier:   94.27 %
```

2. Use Word Embedding Layers with Deep Learning:

After implementing ML models, we are supposed to use word embedding layers with a deep learning model and calculate the accuracy of the model for the text classification task. I have used the glove model for word embedding recommended by the professor. To implement the glove model I have used glove global vectors for word representation and take the reference from Basics of Using Pre-trained GloVe Vectors in Python.

After considering glove data, I have calculated embedding matri and passed it into the embedding initializer to generate a word embedding layer in the deep learning model. Then, I have used adam optimizer with sigmoid activation function with a dropout of 0.4 and 32 dense. As a result, I got the training and validation accuracy as below:



3. Result of testing data:

```
9/9 [==============================] - 0s 4ms/step - loss: 0.3425 - accuracy: 0.8540
[0.3425476849079132, 0.8540399670600891]
```