ROOT2AI TECHNOLOGY PRIVATE LIMITED

Prepared By: Vanshika Nehra

PROBLEM STATEMENT:

META DATA:

1. Text : Contains text from blockchain domain
2. Target : Target class

INFERENCE:

The data contains text sentences classified into 11 categories.

Approach:

First step is to clean the dataset and fill all the rows the are NULL. Data Cleaning requires – a) Removing all the next lines, tabs, punctuation marks, stop words such as "I, am, are, is, etc.". b) Lemmatization i.e., converting each word into a root meaningful word. c) Tokenisation i.e., converting each word into Tokens. d) Converting Target labels into numbers.
Second Step is to apply various classifiers on the cleaned dataset to predict labels. I have used CNN, LSTM, SVM classifiers.

Model Interpretation:

The model contains 4 different classifiers to predict the labels of unseen data. Data is cleaned first then given to the classifiers. For all the classifiers, first val_loss decreases then increases as loss goes down and val_acc remains constant as acc goes up. The models are overfitting. To overcome overfitting I used Dropout layers and also switched to SVC.

Train and Test Score:

CNN with 3 + Dense layers: acc = 0.5059, val_acc = 0.4211 with 20 epochs
CNN with 5 + Dense layers: acc = 0.6235, val_acc = 0.4351 with 15 epochs
LSTM: acc = 0.9579, val_acc = 0.6014 with 15 epochs
SVM: score = 0.2634746403005783
Taking into consideration val_loss also, LSTM performed better than others.

Limitation of the model:

The model with LSTM has val_acc of 60% but val_loss is high which means it is still unable to predict properly.