**Epoch Cores Selection - Round 1**

**Task 2**

**Name: Aparna. S. Kaushik**

**Objective**

The primary objective of this project is to use artificial intelligence to convert handwritten text images into digital text and subsequently perform sentiment analysis on the extracted text. The project will leverage the provided labeled dataset of handwritten alphabets.

**Process Followed:**

1. **Data Filtering and Preprocessing:**

* Convert any column other than the label using convert\_other\_column and the label column using the convert\_label function while loading the dataset.
* The convert\_other\_column checks if it contains an int data type if not replaced by an NAN value, and convert\_label converts the alphabet letter to its int value (0 to 25); otherwise, it converts to NAN value.
* Then, the data containing NAN values was removed.
* Removing duplicate data in the filtered data.
* Splitting the data into training, testing, and validation data (equally distributed label-wise)
* Normalized the input values.

1. **Model Training**

* 2 Hidden layers of 512 units along with dropout with a probability of 0.5.
* Used Early stopping to prevent over-fitting.
* Used Adam optimizer.

1. **Image Preprocessing**

* Did necessary pre-processing accordingly.
* Using contours, identified letters present and spaces present also.
* Discarded some useless contours.
* Line-wise sorting of letters.

1. Model applied on target images and predicted the sentences.
2. Loaded line dataset, created the Naïve-Bayes model, and used it on predicted sentences for sentiment analysis.

**Results:**

Even though the predicted lines are not that accurate, the Naive Bayes Method is able to predict the sentiment of the line correctly.

* I tried adding a token of NOT\_ at the beginning of the word after an occurrence of a negative word in the sentence.
* We should stop adding NOT\_ if we encounter any punctuation or another negative word.
* Since it is a single line, we stop adding only when another negative word is encountered.

Adding this token didn't affect the final results in this case, but this will be a good practice. I believe if any other dataset was given, the result may vary according to the addition of a token. Since it will separate the negative meaning of positive words and it will totally deny the existence of that positive word in the sentence.