CHATBOT USING RNN
GDSC IITK

Problem Statement:

Develop a machine learning model for multi-class text classification using the provided training dataset ("train.csv") containing two columns: text and label. The objective is to create a robust model leveraging NLP libraries and models capable of accurately classifying new text entries into predefined labels. The model's performance will be evaluated on a similar test dataset.

Key Tasks:

- Preprocess the text data, considering tokenization, cleaning, and any necessary normalization steps.
- Utilize NLP libraries (such as NLTK, spaCy, or Hugging Face's Transformers) to engineer relevant features from the text data.
- Experiment with various state-of-the-art NLP models suitable for multi-class classification to determine the most suitable architecture for this task.
- Train the selected model on the provided training dataset ("train.csv") to learn the associations between the text features and their respective labels.
- Evaluate the model's performance using appropriate metrics (e.g., accuracy, precision, recall, F1-score) on a test dataset with similar structure and labels.
- Fine-tune the model if necessary to improve its classification accuracy.

Provide a submission that includes the trained model and instructions on how to use it to predict the labels for new text entries.

Deliverables:

- Trained machine learning model capable of multi-class text classification.
- Evaluation metrics showcasing the model's performance on the test dataset.
- Documentation detailing the methodology, model architecture, preprocessing steps, and instructions for using the model for predictions on new text data.

Submission Guidelines:

Submit the trained model along with a code/script demonstrating how to load the model and predict labels for new text entries. Ensure the code is well-documented and easily understandable for end-users to employ the model for classification tasks.

The success of the submission will be based on the model's accuracy, robustness, and efficiency in classifying unseen text data into the appropriate labels specified in the test dataset.