TASK 2:
Video tagging is the process of labelling videos under various categories. The problem statement demands a novel method to tag videos from social media apps automatically using machine learning techniques which can be used by recommendation systems. We will now have a look in brief about the feasibility, methodology and challenges involved in building this model. FEASIBILITY: While the process of automatically tagging videos is doable, it is highly expensive in general, and it mainly depends on the sampling rate of videos and the dataset that is used for training. The computation time increases for videos of longer duration and videos which have poor metadata content. If the dataset is going to be the watch history(preferences in instagram) of users, then the dataset is required to be enriched with diversified categories. If not, then one has to go for unsupervised learning techniques which normally yields lesser accuracy than supervised learning. As an alternative, I would suggest to go for the parsing of top comments associated with each video in YouTube and Instagram. This involves NLP, which is less expensive than the combination of sampling and image processing. Also the authors(content creators, channels) of that particular video can be considered for the prediction process. The author's previous videos can be taken into account, and the associated comments can be processed for better prediction.

METHODOLOGY: The first step involves pre-processing of key-frames(frames which are not too similar to each other) obtained through sampling over the interval. The obtained frames are then converted to grayscale(to reduce channel size) and are resized by passing them through a set of kernels along with normalization of pixels. By doing these, one can reduce the computation time of the process. Key-frame extraction is quite important, and should not contain redundant frames. For this the color distribution of each frame is analyzed through a histogram and frames are obtained accordingly. Then a Convolutional Neural Network(CNN) is devised, following the VGG16 architecture which is especially used for large scale image processing. However the activation function used in convolutional layers is LeakyRelu, in order to introduce a very small proportion of non-linearity. In image processing techniques, LeakyRelu is deployed the most as it results in lesser dead units than the normal Relu function. The fully connected layer follows sigmoid activation. The kernel sizes for convolutional and MaxPooling is taken as 3x3 and 2x2 respectively, and the stride values are taken as 1x1, as per normal convention. Next follows the Sequential type series of dense hidden layers with Relu activation functions and softmax activation function at the end. At most 5 dense layers are used and dropout layers are used to avoid overfitting. The final dense layer should contain neurons equal to the number of classes(labels). Adam optimizer is used and categorical_crossentropy is used to calculate loss. Based on the obtained accuracy results, the number of epochs, neurons, test_size, batch_size and dropout layers can be altered. This is the overview of the development of the model. CHALLENGES: The key challenge involved in this process is the prediction of more than one tag, for an input video, which is generally not preferred and leads to misinterpretation. For each frame of a video, a prediction is made and for the overall video, two or more classes might have similar prediction counts. In that case, one can remove the classes which had lesser predictions and once again pass the input to the model. Now, for each prediction made for a frame, the count is noted and at the end, the average number of counts is computed and based on the result, one can tag the predicted label. Also, one should be careful in choosing the key-frames and assessing the histogram for each frame. Similar frames might yield incorrect predictions. Another important factor, is the diversity present in the training dataset, and the semantic gap for each value. More diverse values will give predictions with improved accuracy. Other challenges include longer duration of the video which leads to higher computation time and complexity of the process. With this I conclude my answer for the given problem statement.