1) Under weakly supervised pre-training, the existing dataset of Instagram images labelled with their corresponding hashtags are used to train and predict on upto 3.5 Billion images. This model is subsequently fine-tuned on a task specific dataset, which is the final model for which we analyze the performance.

   Under the semi-supervised setup, the paper explores using a teacher-student model: Initially, a relatively smaller setup of labelled images are used to train the “teacher” model. This model is used to predict the labels on a set of unlabelled data, and choose the top ‘K’ ranked images to train a “student” model. The student model is then finally fine-tuned on the initial supervised data used to train the model.

   Although both are using over a billion images, the weakly supervised model is trained on a noisy, yet labelled dataset of a specific set of chosen hashtags, whereas the semi-supervised model is trained on a clean, smaller dataset, which subsequently predicts the outcome on unlabelled data. These outcomes are considered as the truth values, and is used to fine-tune on the initial, cleaner dataset.

2) a) The trained model is **robust** when it comes to noise in the hashtags.
   On the given dataset containing billions of images, filtering the noise is not an easy task. Thus, while training, noisy labels were used. To study the impact of noisy labels on the model, additional noise in the labels were added, in factors of 10%, 25% and 50%, and the model was trained. It showed a remarkable resilience to the addition of such noise- adding 10% showed only a 1% decrease in accuracy, 25% showed a decrease of 2% and 50% only 6%. This is a stunning result, and shows that at the scale of a billion images, the models are inherently noise resistant.

   A reason for the above may be that the noise added during pre training was masked during the process of fine tuning.

   b) The inherent hashtag data does not appear to have any distribution, whereas the hashtags seem to be governed by the Zipfian distribution, which follows the Zipf’s law of inverse rankings. Thus, we choose to sample images along a Zipfian distribution, this reduces the overall training risk as the model tries to represent the actual distribution, thus reducing the difference between the true and empirical risks.

   Supporting experiments show that a square root sampling or a uniform sampling increases the model accuracy by upto 5%.

3) a) The teacher model is typically a huge neural network, the aim of which is to reduce the empirical risk of the model. The amount of computation is not a factor for this network. The student model is typically a small network, with inferencing computation in mind.

The aim of using both these models is an attempt to achieve a high amount of accuracy (reproduce the output of the bigger model) despite using a small model- ie we would like the student model to perform as well as the teacher model, despite being a much smaller in size.

The way it achieves this requirement is by allowing the teacher model to make predictions, and use this predicted data to train the smaller model. Thus, as the training loss reduces, the smaller model mimics the performance of the bigger model. The student model can then be fine tuned for a specific task, and provides an accuracy similar to the bigger one.

b) The training and testing datasets may not come from identical distributions since the test set on which we predict are unlabelled images. The probability that the test set contains labels which are not in the training set is high, and underrepresentation of certain classes in the training set is also possible. These factors can have a huge negative impact on the predictions- they essentially act as noise in the dataset.

To overcome this difficulty, instead of using all the unlabelled examples during prediction using the teacher model, we only select top K examples- the ones which our model predicts with the highest of confidence. This is obtained by using a softmax predictor on the training set, and choosing the top K ranked images to represent each label.

Given that we have K examples for each class, the next step is to use only classes which we know, with high confidence for each image. Each image can be associated with a number of classes, but the probability that it is associated with a large number of classes is very low. The choise of P>1 is governed by the fact that some classes may be under-represented, and may thus have just very few images pertaining to it, and this associating these images with multiple labels result in our model never predicting these classes. However, the higher the value of P, better the balance in the distribution.

c)
Initially, a supervised dataset D is used to train the teacher model.

Next, the set of unlabelled examples U is fed to a softmax predictor. For each target class, K top ranked images are selected- ie K images with the highest values of softax outputs is defined to belong to this class. At this stage, each image may pertain to one or more classes. To balance this distribution, each image is limited to its top P classes. The choice of P used is 10 for the experiments in the paper.

Taking into account K and P, the resultant dataset labelled \hat{D} is used to train the student model.

Finally, the student model is then fed the original, labelled, clean dataset for finetuning.

d) The parameter K can be used to explain Figure 5- the variation in accuracy. From left to right in the graph, the value of K increases from 0 to 16K.

The initial increase in accuracy can be attributed to the fact that as K increases, the number of images associated with each class increases. Thus, there are more images for the model to learn to predict for a particular class. At K= 8000 - 16000, the model seems to reach its peak performance, where each target class has a good number of “correct” training examples to train on.

Beyond 16K, the performance starts to reduce. This can be attributed to the fact that for high values of K, we are adding to the noise in the dataset - ie for each target class, we include images which, by prediction, belong in the lower confidence bounds. Thus, labels which may not belong to our target class get labelled incorrectly. Using these examples to train the student model will result in wrong predictions.