PROPOSAL

- **PROJECT:** Text Sentiment Analysis

- **PROBLEM DESCRIPTION:**

The current model focuses more on a binary classification of sentiment(either positive or negative) based on the input sentence and a sentiment score generated thereof. But we are at a disadvantage of capturing subtle intricacies in emotions of different sentences. For example: the two sentences: I was surprised that the match was scheduled today and I was surprised by the fact that they did a miracle.These two sentences showcase a different range of positive emotions which when classified as only positive for both cases won't justify the emotion clearly.In addition to that the model will predict poorly on contemporary writing style(mostly social media texts).

- **MY PROPOSAL:** A hybrid fine grained embedding based(A pre-trained model+A customized transformer model+ fine tuning) NLP model trained on a more elaborate data set.

- **DETAILS:** The idea is to make train the model first on a bidirectional transformer encoder model(Preferably BERT) with the help of a more distinguishing dataset-The SST(Stanford Sentiment Treebank) and then add a layer of a customized layerof transformer and fine tuning to enhance the model's predicting power.
  It will consist of three phases:
  1. **PREPARING THE DATASET AND TEXT PREPROCESSING:**
     The dataset to be used is the SST dataset. The main advantage of using this dataset is that it is more extensive relating to a range of emotions between positive and negative. Furthermore ,there will be an addition of a set of contemporary sentences mostly used in social media to have an added variance in te dataset. After that the normal text preprocessing steps are to be conducted in a graded manner(conversion of the dataset to csv,removal of stop words and redundancies etc.).This will be done using the popular Regular expression library.

2. **TRAINING THE MODEL:** The model will be trained initially on a BERT based model with some initial fine tunings.Since the model is

powerful(BERT) it will capture most of the words and their synonyms easily.

3. **CUSTOMIZED ADDITION ,FINETUNING AND EVALUATION:**
   Here there will be an addition of a customized transformer layer using tensorflow in a graded manner which will include: 1.Mini-batch division
   2. normalization over the batches
   3. Self attention
   4. Dropout
   5. A neural network with non-linearity(preferably ReLu)
   6. And two masks: a.attention masks and b.padding mask

This will help to enhance the model accuracy furthermore which may shoot up to a range of 50-60 percent. The evaluation can be done using a confusion matrix.

- TECH STACK TO BE USED:
  1. PYTHON
  2. TENSORFLOW
  3. REGEX
  4. NLTK
  5. PANDAS
  6. NUMPY

- TIMELINE:

1. 7th-10th december: Familiarize with the code and the community, the version control system, the documentation and nitty gritties of the tech stack to be used throught the process.

2. 10th -20th december: Execute the above mentioned stages of the project in phases with the help of the mentor's guidence.

3. Mid term assesment: Collaborate with the mentor to evalauate the code and make changes if necessary

4. 23rd -5th january: proceed to phase 2 after first evaluation is completed.