Progress report :- ML Project

(1) Data Pre-processing : what data have you collected so far and what preprocessing have you done?

"We acquired our dataset from two sources,
Data1
Data2

Preprocessing steps used : - Using neattext module

1. Lowercasing: Convert all text to lowercase. This ensures uniformity and reduces the complexity of the data.

2. Removing Punctuation: Remove any special characters and punctuation marks from the tweets as they don't usually contribute significantly to emotion detection.

3. Removing Stopwords: Remove common stop words (e.g., 'and', 'is', 'in') that do not carry significant meaning and are often noise in text data.

4. Handling Emoticons and Special Characters: Emoticons and special characters might carry emotional meaning. You can choose to keep, remove, or replace them based on the context of your analysis.

5. Handling URLs and User Mentions: Remove or replace URLs and user mentions (starting with '@') as they usually don't contribute to emotion detection.

6. Handling Numbers: Decide whether to keep numbers, convert them to words, or remove them based on the context of your analysis.

(2) Training with the basic model, validation, and completion of the data pipeline.Which models did you use, what training/validation accuracy have you achieved?
Is your data pipeline completed?

We used 3 different models :-

1. Logistic regression
   Score after using Logistic regression : `0.6054799770070894`

2. SVM
   Score using SVM : 0.5966660279747078

3. Random Forest
   Score using random forest : 0.558919333205595

For the current models the Pipeline is complete from taking the test cases , data preprocessing Countvectorising and model being used to predict.

(3) Identification of the exact tasks you want to complete for the final submission. What challenges you are facing and how you plan to address them. What will be your final deliverables?

Exact task we want to complete for final submission is to make a neural network (using RNN or LSTM) which takes a text input and returns the emotion with which the text was written and if neural network is completed on time then we will be making a web app on which some comments can be posted . The comments will be posted after being checked by the ML program and a reaction emoji will be automatically generated based on the results of the ML program.

The challenges faced by us till now is that the ML program is trying to detect emotion without knowing the complete meaning of the sentence. Which is why we will be using RNN and LSTM for further improvement.

Another challenge faced by us is that the dataset we are using is not uniformly distributed and because of this some emotions are not getting sufficient data to train the program. We are looking for more data so we can train the model correctly. If we don't find the data we need we will merge some emotions like anger and disgust , shame and fear to make it more uniform.

Each group member needs to identify what they have done so far.

1. **Aditya Singh :** Data preprocessing and Linear Regression model
2. **Lalit Gour :** Random Forest model and Build Pipelines
3. **Kunal Singla :** Making all the functions and SVM model