Problem statement:

We are given a dataset which has the data of surveys for 3 different wheels, and emotions were recorded for each user. So the dataset has user id, clip id, time, emotions and intensity. Each wheel has test data of several different clips which were played during the survey. Emotion was recorded for each clip and the time corresponding to the click from the start of the clip was also recorded. After going through the data we decided to predict the emotions for a particular song for any user on the basis of given data. Further we decided to generalize it by predicting the emotion for any song and any user.

Approach:

Given different clips , we extracted the frequency spectrum using a third party software Audacity. We tried to study the basic features (frequency and amplitude) of a clip which can be used to determine the parameters responsible for the emotion in a given clip. Plotting the variation of frequency and amplitude with time will give a broad view of analyzing the dataset. So a dataset can be made having frequency and amplitude as features which is further depending on time and our emotion as our outcome.
After reading few research papers and the work carried out in this field gave a rudimentary idea of considering ragas and notes as features as well. Ragas are combinations of notes in a particular sequence which can be further classified as a range of frequency.

Implementation and Problem Faced:

After getting the frequency spectrum from audacity, we exported the data of frequency spectrum into a excel file. This file consists of total of 256 values of frequency and amplitude. So we decided to create a time interval for each clip on the basis of their corresponding length. This time interval is determined by dividing the length of the clip with 256 so as to get a uniform data set. Now we have our frequency and amplitude varying with time which is our feature of the training data set. This features were used to determine the emotions for a given clip for any user as discussed before.
But here the problem is that the frequency spectrum is independent of time. The frequency spectrum is generally the fast fourier transform of the data. Now fourier transform basically gives the variation of dominant frequency with amplitude which is independent of time. So our approach of dividing the 256 values of frequency and amplitude into time intervals is fundamentally wrong.
So, our another approach was to extract the frequency spectrum for a specified time interval and getting various frequency spectrums for each clip. Then further training this frequency spectrum with the emotions, we can predict the emotions. Here again the problem is for a given time interval how to choose a particular value of frequency and amplitude which can be used for training, since frequency spectrum in itself is a distribution and cannot be directly used as a single feature.

So basically let us say that we have a clip of 10 seconds and we break it into 1 second each and extract the frequency spectrum of all ten seconds separately and the problem is that these spectrums are distributions and cannot be considered as a single feature.

We were also thinking of adding additional features so that we can get a better accuracy results but found that almost every feature we explored were in some way or another were dependent on frequency and amplitude, we are currently kind of stuck at this step.

We were also thinking of using MIRTOOLBOX (a matlab based toolkit) for feature extraction from the audio.