Gowtham Vuppala
Task 1 Report
StudentID: 001270914

NOTE: For the purpose of labelling using python, I changed database_5.xlsx to database_5.csv(UTF-8 encoded), because when I'm trying to load database_5.xlsx using numpy or pandas, I kept on getting errors where it keeps saying that the database_5.xlsx has characters that UTF-8 can't encode. So, I changed the extension to .csv without any compromise in the data.

Report:

1) For labelling, I wrote a python program which labels as 1 if there is a bad word in the comment, else labels as 0. To make sure it labels correct, I used a bad words list with almost 500 words. Due to not manually labelling, there are cases where it labels as 1 when there is a bad word in a comment, but it isn't really a negative comment. Except for such cases, my program labels accurately for almost all other comments.

2) For Role Categorizing, I couldn't come up with a perfect solution. So, I labelled every comment with a label 1 as a 'bully'(since most of them are bullies) and every other comment role as 'other'.