Data Collection and Preprocessing Phase

Date	15 March 2024
Team ID	739724
Project Title	Analysis of amazon cell phone reviews
Maximum Marks	6 Marks

Preprocessing Template

The images will be preprocessed by Text preprocessing , import libraries and text cleaning.

Section	**Description**
Text Preprocessing	Data analysing , data cleaning , filling the null values.
Import libraries	Taking required libraries for cleaning and preprossesing the data.
Text cleaning	Changing the size of the font and filling the empty comments
Data Preprocessing Code Screenshots	
Text Preprocessing	-

Import libraries	```#import natural language toolkit import nltk #import stopwords library to remove stopwords from nltk.corpus import stopwords #library used for stem the words from nltk.stem.porter import PorterStemmer #create an object for stemming ps = PorterStemmer() #library used for stem the words from nltk.stem import WordNetLemmatizer #create an object for wordnet lemmatizer wordnet=WordNetLemmatizer()```
Text cleaning-1	```# Initialize empty array to append clean text corpus=[] # no of rows to clean for i in range(len(x)): #replacing punctuations and numbers using re library temp=re.sub('[^a-zA-Z]',' ',x[i]) # convert all text to lower cases temp=temp.lower() # split to array(default delimiter is " ") temp=temp.split() # creating WordNetLemmatizer object to take main lemma of each word wordnet = WordNetLemmatizer() #loop for leammatization each word in string array at ith row temp=[wordnet.lemmatize(word) for word in temp if not word in set(stopwords.words('english'))]```
Text cleaning-2	```[ ] #creating bag of word model from sklearn.feature_extraction.text import CountVectorizer #To extract max 2000 feature, "max_features" is attribute to #experiment with to get better results cv=CountVectorizer(max_features= 2000) #z contains vectorized data (independent variable) z=cv.fit_transform(corpus).toarray()``` Save the Bag of word model ```[ ] import pickle pickle.dump(cv,open('count_vec.pkl','wb'))``` ```from sklearn.model_selection import train_test_split x_train,x_test,y_train,y_test=train_test_split(z,y,test_size=0.3)```