Importing the libraries

For numerical calculations and data handling

```
import numpy as numpy
import pandas as pd
```

For visualization of data in the project

```
import matplotlib
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
```

```
import sklearn
from sklearn.utils import shuffle
from sklearn.feature_extraction.text import TfidfVectorizer
```

NLP preprocessing libraries

```
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
```

```
import re
import random
import warnings
warnings.filterwarnings(action='ignore', category=UserWarning, module='gensim')
import gensim
```

```
from collections import Counter
import unicodedata as udata
import string
```

Checking the versions

```
print(sklearn.__version__)
print(matplotlib.__version__)
print(numpy.__version__)
print(pd.__version__)
print(nltk.__version__)
```

```
0.21.2
3.1.0
1.16.4
0.24.2
3.4.4
```

Reading the csv files

```
trainSet = pd.read_csv("train.csv", encoding='latin-1', header=None)
testSet = pd.read_csv("test.csv", encoding='latin-1', header=None)
trainSet
```

	0	1	2
0	id	label	tweet
1	1	0	@user when a father is dysfunctional and is s...
2	2	0	@user @user thanks for #lyft credit i can't us...
3	3	0	bihday your majesty
4	4	0	#model i love u take with u all the time in ...
5	5	0	factsguide: society now #motivation
6	6	0	[2/2] huge fan fare and big talking before the...
7	7	0	@user camping tomorrow @user @user @user @use...
8	8	0	the next school year is the year for exams.Ã°Â...
9	9	0	we won!!! love the land!!! #allin #cavs #champ...
10	10	0	@user @user welcome here ! i'm it's so #gr...
11	11	0	Ã¢Â□Â□ #ireland consumer price index (mom) cl...
12	12	0	we are so selfish. #orlando #standwithorlando ...
13	13	0	i get to see my daddy today!! #80days #getti...
14	14	1	@user #cnn calls #michigan middle school 'buil...
15	15	1	no comment! in #australia #opkillingbay #se...
16	16	0	ouch...junior is angryÃ°Â□Â□Â□#got7 #junior #y...
17	17	0	i am thankful for having a paner. #thankful #p...
18	18	1	retweet if you agree!
19	19	0	its #friday! Ã°Â□Â□Â□ smiles all around via ig...
20	20	0	as we all know, essential oils are not made of...
21	21	0	#euro2016 people blaming ha for conceded goal ...
22	22	0	sad little dude.. #badday #coneofshame #cats...
23	23	0	product of the day: happy man #wine tool who'...
24	24	1	@user @user lumpy says i am a . prove it lumpy.
25	25	0	@user #tgif #ff to my #gamedev #indiedev #i...
26	26	0	beautiful sign by vendor 80 for $45.00!! #upsi...
27	27	0	@user all #smiles when #media is !! Ã°Â□Â□Â...
28	28	0	we had a great panel on the mediatization of t...
29	29	0	happy father's day @user Ã°Â□Â□Â□Ã°Â□Â□Â□Ã°Â□Â...
...	...	...	...
31933	31933	0	@user thanks gemma
31934	31934	1	@user judd is a &amp; #homophobic #freemilo #...
31935	31935	1	lady banned from kentucky mall. @user #jcpenn...
31936	31936	0	ugh i'm trying to enjoy my happy hour drink &a...
31937	31937	0	want to know how to live a life? do more thi...
31938	31938	0	love island Ã°Â□Â□Â□
31939	31939	0	my fav actor #vijaysethupathi ! my fav actress...
31940	31940	0	whew Ã°Â□Â□Â□ it's a productive and #friday!!!
31941	31941	0	@user she's finally here! @user
31942	31942	0	passed first year of uni #yay #love #pass #uni...
31943	31943	0	this week is flying by #humpday - #wednesday...
31944	31944	0	@user modeling photoshoot this friday yay #mo...
31945	31945	0	you're surrounded by people who love you (even...
31946	31946	0	feel like... Ã°Â□Â□Â□Ã°Â□Â□Â¶Ã°Â□Â□Â□ #dog #su...

31947	31947	1	@user omfg i'm offended! i'm a mailbox and i'...
31948	31948	1	@user @user you don't have the balls to hashta...
31949	31949	1	makes you ask yourself, who am i? then am i a...
31950	31950	0	hear one of my new songs! don't go - katie ell...
31951	31951	0	@user you can try to 'tail' us to stop, 'butt...
31952	31952	0	i've just posted a new blog: #secondlife #lone...
31953	31953	0	@user you went too far with @user
31954	31954	0	good morning #instagram #shower #water #berlin...
31955	31955	0	#holiday bull up: you will dominate your bul...
31956	31956	0	less than 2 weeks Ã°Â□Â□Â□Ã°Â□Â□Â□Ã°Â□Â□Â¼Ã°Â□...
31957	31957	0	off fishing tomorrow @user carnt wait first ti...
31958	31958	0	ate @user isz that youuu?Ã°Â□Â□Â□Ã°Â□Â□Â□Ã°Â□Â...
31959	31959	0	to see nina turner on the airwaves trying to...
31960	31960	0	listening to sad songs on a monday morning otw...
31961	31961	1	@user #sikh #temple vandalised in in #calgary,...
31962	31962	0	thank you @user for you follow

31963 rows × 3 columns

Just removing the first row as it is of no use

```
trainSet = trainSet.drop([0], axis=0)
trainSet
```

	0	1	2
1	1	0	@user when a father is dysfunctional and is s...
2	2	0	@user @user thanks for #lyft credit i can't us...
3	3	0	bihday your majesty
4	4	0	#model i love u take with u all the time in ...
5	5	0	factsguide: society now #motivation
6	6	0	[2/2] huge fan fare and big talking before the...
7	7	0	@user camping tomorrow @user @user @user @use...
8	8	0	the next school year is the year for exams.Ã°Â...
9	9	0	we won!!! love the land!!! #allin #cavs #champ...
10	10	0	@user @user welcome here ! i'm it's so #gr...
11	11	0	Ã¢Â□Â□ #ireland consumer price index (mom) cl...
12	12	0	we are so selfish. #orlando #standwithorlando ...
13	13	0	i get to see my daddy today!! #80days #getti...
14	14	1	@user #cnn calls #michigan middle school 'buil...
15	15	1	no comment! in #australia #opkillingbay #se...
16	16	0	ouch...junior is angryÃ°Â□Â□Â□#got7 #junior #y...
17	17	0	i am thankful for having a paner. #thankful #p...
18	18	1	retweet if you agree!
19	19	0	its #friday! Ã°Â□Â□Â□ smiles all around via ig...
20	20	0	as we all know, essential oils are not made of...
21	21	0	#euro2016 people blaming ha for conceded goal ...
22	22	0	sad little dude.. #badday #coneofshame #cats...
23	23	0	product of the day: happy man #wine tool who'...
24	24	1	@user @user lumpy says i am a . prove it lumpy.
25	25	0	@user #tgif #ff to my #gamedev #indiedev #i...
26	26	0	beautiful sign by vendor 80 for $45.00!! #upsi...
27	27	0	@user all #smiles when #media is !! Ã°Â□Â□Â...
28	28	0	we had a great panel on the mediatization of t...
29	29	0	happy father's day @user Ã°Â□Â□Â□Ã°Â□Â□Â□Ã°Â□Â...
30	30	0	50 people went to nightclub to have a good nig...
...	...	...	...
31933	31933	0	@user thanks gemma
31934	31934	1	@user judd is a &amp; #homophobic #freemilo #...
31935	31935	1	lady banned from kentucky mall. @user #jcpenn...
31936	31936	0	ugh i'm trying to enjoy my happy hour drink &a...
31937	31937	0	want to know how to live a life? do more thi...
31938	31938	0	love island Ã°Â□Â□Â□
31939	31939	0	my fav actor #vijaysethupathi ! my fav actress...
31940	31940	0	whew Ã°Â□Â□Â□ it's a productive and #friday!!!
31941	31941	0	@user she's finally here! @user
31942	31942	0	passed first year of uni #yay #love #pass #uni...
31943	31943	0	this week is flying by #humpday - #wednesday...
31944	31944	0	@user modeling photoshoot this friday yay #mo...
31945	31945	0	you're surrounded by people who love you (even...
31946	31946	0	feel like... Ã°Â□Â□Â□Ã°Â□Â□Â¶Ã°Â□Â□Â□ #dog #su...

31947	31947	1	@user omfg i'm offended! i'm a mailbox and i'...
31948	31948	1	@user @user you don't have the balls to hashta...
31949	31949	1	makes you ask yourself, who am i? then am i a...
31950	31950	0	hear one of my new songs! don't go - katie ell...
31951	31951	0	@user you can try to 'tail' us to stop, 'butt...
31952	31952	0	i've just posted a new blog: #secondlife #lone...
31953	31953	0	@user you went too far with @user
31954	31954	0	good morning #instagram #shower #water #berlin...
31955	31955	0	#holiday bull up: you will dominate your bul...
31956	31956	0	less than 2 weeks Ã°Â□Â□Â□Ã°Â□Â□Â□Ã°Â□Â□Â¼Ã°Â□...
31957	31957	0	off fishing tomorrow @user carnt wait first ti...
31958	31958	0	ate @user isz that youuu?Ã°Â□Â□Â□Ã°Â□Â□Â□Ã°Â□Â...
31959	31959	0	to see nina turner on the airwaves trying to...
31960	31960	0	listening to sad songs on a monday morning otw...
31961	31961	1	@user #sikh #temple vandalised in in #calgary,...
31962	31962	0	thank you @user for you follow

31962 rows × 3 columns

Shuffling the data in the data frame

```
trainSet = trainSet.sample(frac=1).reset_index(drop=True)
```

Assigning names to the columns

```
trainSet.columns = ["id", "sentiment", "tweet"]
```

```
trainSet.columns
```

```
Index(['id', 'sentiment', 'tweet'], dtype='object')
```

Checking null values in the dataset, here we are counting null values in each column in the dataset

```
trainSet.isnull().sum()
```

```
id           0
sentiment    0
tweet        0
dtype: int64
```

Checking the duplicates values and counting duplicates in the data set

```
trainSet.duplicated().sum()
```

```
0
```

Get the first five rows from the dataset

```
trainSet.head(5)
```

	id	sentiment	tweet
0	5755	0	a superb day for @user golf day at @user cours...

drop some unwanted column from the dataframe

```
trainSet = trainSet.drop(["id"], axis = 1)
```

4 238 0 can't wait for carowinds! ready to see @user ...

```
trainSet.head(5)
```

	sentiment	tweet
0	0	a superb day for @user golf day at @user cours...
1	0	@user ahhhh might have guessed #euro2016
2	0	@user spent my entire day trying to so new con...
3	1	@user @user lock the gate! #pamgelleheracist ...
4	0	can't wait for carowinds! ready to see @user ...

count the number of sentiments with respect to their tweet (0 stands for positive tweet and 1 stands for negative tweet)

```
trainSet.sentiment.value_counts()
```

```
0    29720
1     2242
Name: sentiment, dtype: int64
```

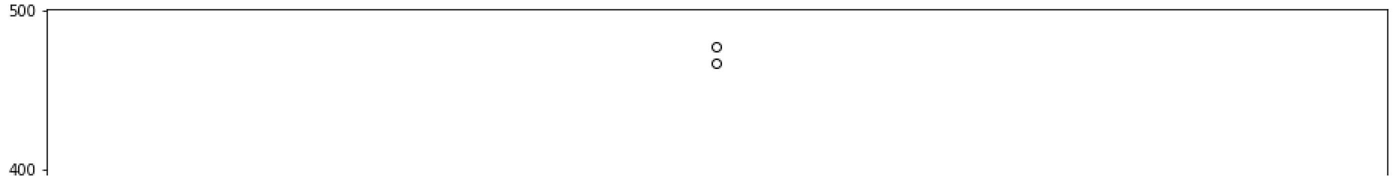
Cleaning data

add new column pre_clean_len to dataframe which is length of each tweet

```
trainSet['pre_clean_len'] = [len(t) for t in trainSet.tweet]
```

Finding outliers using Box plot using pre_clean_len column

```
plt.boxplot(trainSet.pre_clean_len)
fig = plt.gcf()
fig.set_size_inches(16,10)
plt.show()
```



As there are outliers, after preprocessing, we will again test for outliers to see if we got rid of them

```
print(trainSet.shape)
```

```
(31962, 3)
```

Cleaning Operations

Importing beautiful soup

remove @ mentions from tweets

remove URLs from tweets

converting words like isn't to is not

get only text from the tweets

remove utf-8-sig code

converting all into lower case

will replace non-alphabetic characters by space

Word Punct Tokenize and only consider words whose length is greater than 1

join the words

```
import re
from bs4 import BeautifulSoup
from nltk.tokenize import WordPunctTokenizer
tok = WordPunctTokenizer()

pat1 = r'@[A-Za-z0-9_]+'        # remove @ mentions from tweets
pat2 = r'https?://[^ ]+'        # remove URLs from tweets
combined_pat = r'|'.join((pat1, pat2)) #addition of pat1 and pat2
www_pat = r'www.[^ ]+'         # remove URLs from tweets
negations_dic = {"isn't":"is not", "aren't":"are not", "wasn't":"was not", "weren't":"were not",   # converting words like isn't to
                "haven't":"have not","hasn't":"has not","hadn't":"had not","won't":"will not",
                "wouldn't":"would not", "don't":"do not", "doesn't":"does not","didn't":"did not",
                "can't":"can not","couldn't":"could not","shouldn't":"should not","mightn't":"might not",
                "mustn't":"must not"}
neg_pattern = re.compile(r'\b(' + '|'.join(negations_dic.keys()) + r')\b')

def tweet_cleaner(text):  # define tweet_cleaner function to clean the tweets
    soup = BeautifulSoup(text, 'lxml')    # create beautiful soup object
    souped = soup.get_text()   # get only text from the tweets
    try:
        bom_removed = souped.decode("utf-8-sig").replace(u"\ufffd", "?")    # remove utf-8-sig code
    except:
        bom_removed = souped
    stripped = re.sub(combined_pat, '', bom_removed) # calling combined_pat
    stripped = re.sub(www_pat, '', stripped) #remove URLs
    lower_case = stripped.lower()      # converting all into lower case
    neg_handled = neg_pattern.sub(lambda x: negations_dic[x.group()], lower_case) # converting words like isn't to is not
    letters_only = re.sub("[^a-zA-Z]", " ", neg_handled)       # will replace # by space
    words = [x for x  in tok.tokenize(letters_only) if len(x) > 1] # Word Punct Tokenize and only consider words whose length is gre
    return (" ".join(words)).strip() # join the words
```

```
limit=31962
clean_tweet_texts = [] # initialize list
for i in range(0,limit): # batch process almost 32000 tweets
    clean_tweet_texts.append(tweet_cleaner(trainSet['tweet'][i]))  # call tweet_cleaner function and pass parameter as all the tweet
```

clean_tweet_texts

```
nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\User\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
True
```

tokenize word in clean_tweet_texts and append it to word_tokens list

```
word_tokens = [] # initialize list for tokens
for word in clean_tweet_texts:  # for each word in clean_tweet_texts
    word_tokens.append(word_tokenize(word)) #tokenize word in clean_tweet_texts and append it to word_tokens list
```

Lemmatizing

```
nltk.download('wordnet')
```

```
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\User\AppData\Roaming\nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
True
```

```
df1 = [] # initialize list df1 to store words after lemmatization
from nltk.stem import WordNetLemmatizer # import WordNetLemmatizer from nltk.stem
lemmatizer = WordNetLemmatizer() # create an object of WordNetLemmatizer
for l in word_tokens: # for loop for every tokens in word_token
    b = [lemmatizer.lemmatize(q) for q in l] #for every tokens in word_token lemmatize word and give it to b
    df1.append(b) #append b to list df1
```

df

```
clean_df1 =[] # initialize list clean_df1 to join word tokens after lemmatization
for c in df1:  # for loop for each list in df1
    a = " ".join(c) # join words in list with space in between and give it to a
    clean_df1.append(a) # append a to clean_df1
```

clean_df

convert clean_tweet_texts into dataframe and name it as clean_df

```
clean_df = pd.DataFrame(clean_df1,columns=['text']) # convert clean_tweet_texts into dataframe and name it as clean_df
#clean_df['target'] = df.sentiment[:10000] # from earlier dataframe get the sentiments of each tweet and make a new column in clean_
#clean_df
```

```
clean_df['clean_len'] = [len(t) for t in clean_df.text] # Again make a new coloumn in the dataframe and name it as clean_len which
```

```
clean_df[clean_df.clean_len > 140].head(10) # again check if any tweet is more than 140 characters
```

	text	clean_len

No outliers anymore

```
target2 = [] # initialize list
for i in range(0,limit): # batch process 32K tweets
    target2.append(trainSet['sentiment'][i])
clean_df['target']=target2
clean_df.head()
```

	text	clean_len	target
0	superb day for golf day at course in fine nick...	66	0
1	ahhhh might have guessed euro	29	0
2	spent my entire day trying to so new contract ...	118	0
3	lock the gate pamgelleheracist mmiw mmiwg	41	1
4	can not wait for carowinds ready to see and ti...	82	0

```
X = clean_df.text # get all the text in x variable
y = clean_df.target # get all the sentiments into y variable
print(X.shape) #print shape of x
print(y.shape) # print shape of y
from collections import Counter
print(set(y)) # equals to list(set(words))
print(Counter(y).values()) #
```

```
(31962,)
(31962,)
{'1', '0'}
dict_values([29720, 2242])
```

Remember 1 is for racist/sexist tweets and 0 is for non-racist/non-sexist tweets

perform train and test split

X_train is the tweets of training data, X_test is the testing tweets which we have to predict, y_train is the sentiments of tweets in the traing data and y_test is the sentiments of the tweets which we will use to measure the accuracy of the model

```
from sklearn.model_selection  import train_test_split #from sklearn.model_selection import train_test_split to split the data into t
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.20, random_state= 1) # split the data into traing and testin
```

Get Tf-idf object and save it as vect. We can select features from here we just have simply change

the ngram range to change the features also we can remove stop words over here with the help of stop parameter

```
vect = TfidfVectorizer(analyzer = "word", ngram_range=(1,3))
```

fit or training data tweets to vect

transform our training data tweets

```
vect.fit(X_train)
X_train_dtm = vect.transform(X_train)
```

transform our testing data tweets

```
X_test_dtm = vect.transform(X_test)
```

Naive Bayes

```
from sklearn.naive_bayes import MultinomialNB # import Multinomial Naive Bayes model from sklearn.naive_bayes
nb = MultinomialNB(alpha = 10) # get object of Multinomial naive bayes model with alpha parameter = 10
```

```
nb.fit(X_train_dtm, y_train)# fit our both training data tweets as well as their sentiments to the multinomial naive bayes model
```

```
MultinomialNB(alpha=10, class_prior=None, fit_prior=True)
```

```
from sklearn.model_selection import cross_val_score  # import cross_val_score from sklear.model_selection
accuracies = cross_val_score(estimator = nb, X = X_train_dtm, y = y_train, cv = 10) # do K- fold cross validation on our traing data
accuracies.mean() # measure the mean accuray of 10 fold cross validation
```

```
0.9297197079701955
```

predict the sentiments of testing data tweets

```
y_pred_nb = nb.predict(X_test_dtm)
```

measure the accuracy of our model on the testing data

```
from sklearn import metrics # import metrics from sklearn
metrics.accuracy_score(y_test, y_pred_nb)
```

```
0.9357109338338808
```

plot the confusion matrix between our predicted sentiments and the original testing data sentiments

```
from sklearn.metrics import confusion_matrix # import confusion matrix from the sklearn.metrics
confusion_matrix(y_test, y_pred_nb)
```

```
array([[5967,    0],
       [ 411,   15]], dtype=int64)
```

Nearest Neighbour

```
from sklearn.neighbors import KNeighborsClassifier
clf_knn = KNeighborsClassifier(n_neighbors=5`)
```

```
clf_knn.fit(X_train_dtm, y_train)
```

```
KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
                     metric_params=None, n_jobs=None, n_neighbors=5, p=2,
                     weights='uniform')
```

```
from sklearn.model_selection import cross_val_score
accuracies = cross_val_score(estimator = clf_knn, X = X_train_dtm, y = y_train, cv = 10)
accuracies.mean()
```

```
0.9438773837670503
```

```
y_pred_knn = clf_knn.predict(X_test_dtm)
```

```
from sklearn import metrics
metrics.accuracy_score(y_test, y_pred_knn)
```

```
0.9485374628499922
```

```
from sklearn.metrics import confusion_matrix
confusion_matrix(y_test, y_pred_knn)
```

```
array([[5959,    8],
       [ 321,  105]], dtype=int64)
```

Applying the models on test.csv

```
testSet = testSet.drop([0], axis=0)
```

```
testSet
```

	0	1
1	31963	#studiolife #aislife #requires #passion #dedic...
2	31964	@user #white #supremacists want everyone to s...
3	31965	safe ways to heal your #acne!! #altwaystohe...
4	31966	is the hp and the cursed child book up for res...
5	31967	3rd #bihday to my amazing, hilarious #nephew...
6	31968	choose to be :) #momtips
7	31969	something inside me dies Ã°Â□Â□Â¦Ã°Â□Â□Â¿Ã¢Â□Â...
8	31970	#finished#tattoo#inked#ink#loveitÃ¢Â□Â¤Ã¯Â¸Â□ ...
9	31971	@user @user @user i will never understand why...
10	31972	#delicious #food #lovelife #capetown mannaep...
11	31973	1000dayswasted - narcosis infinite ep.. make m...
12	31974	one of the world's greatest spoing events #l...
13	31975	half way through the website now and #allgoing...
14	31976	good food, good life , #enjoy and Ã°Â□Â□Â□Ã°...
15	31977	i'll stand behind this #guncontrolplease #se...
16	31978	i ate,i ate and i ate...Ã°Â□Â□Â□Ã°Â□Â□Â□ #ja...
17	31979	@user got my @user limited edition rain or sh...
18	31980	&amp; #love &amp; #hugs &amp; #kisses too! how...
19	31981	Ã°Â□Â□ÂÃ°Â□Â□Â□Ã°Â□Â□Â□ #girls #sun #fave @...
20	31982	thought factory: bbc neutrality on right wing ...
21	31983	hey guys tommorow is the last day of my exams ...
22	31984	@user @user @user #levyrroni #recuerdos mem...
23	31985	my mind is like Ã°Â□Â□Â□Ã°Â□Â□Â□Ã°Â□Â□Â½Ã°Â□Â□...
24	31986	never been this down on myself in my entire li...
25	31987	check twitterww - trends: "trending worldwide...
26	31988	i thought i saw a mermaid!!! #ceegee #smcr ...
27	31989	chick gets fucked hottest naked lady
28	31990	happy bday lucyÃ¢Â□Â¨Ã¢Â□Â¨Ã°Â□Â□Â□ xoxo #love...
29	31991	haroldfriday have a weekend filled with sunbe...
30	31992	@user @user tried that! but nothing - will try...
...	...	...
17168	49130	people do anything for fucking attention nowad...
17169	49131	creative bubble got burst Ã°Â□Â□Â¢ looking for...
17170	49132	tomorrow is gonna be a big day! we are going t...
17171	49133	i am thankful for baby giggles. #thankful #pos...
17172	49134	#model i love u take with u all the time in ...
17173	49135	in life u will grow to learn some pple will wo...
17174	49136	Ã°Â□Â□Â□i was the storm,you were the rain. tog...
17175	49137	lovelgq - broken ep via #rnb #love #heabrok...
17176	49138	spread love not hateÃ¢Â□Â¤Ã¯Â¸Â□Ã°Â□Â□Â□Ã°Â□Â□...
17177	49139	@user @user are the most racist pay ever!!!!!
17178	49140	i am thankful for children. #thankful #positiv...
17179	49141	liverpool Ã¢Â□Â¤Ã¯Â¸Â□Ã°Â□Â□Â¬Ã°Â□Â□Â§ #walk #...
17180	49142	#bakersfield rooster simulation: i want to c...
17181	49143	por do sol Ã³Â¾Â¯Â□Ã¢Â Â¤Ã¯Â¸Â÷#instagood #bea...