b4fkbcxr1

December 30, 2024

1 Assignment-1

Cohort 11 - PGP in AI/ML

C5 - Text Mining

Assignment - Sentiment Analysis Using Naive Bayes Perform Text Classification on the data. The tweets related to coronavirus have been pulled from Twitter, and manual tagging has been done.

You might use some of the References given below:

1. Sklearn Pipeline
2. Sklearn GridSearchCV
3. ML Pipeline with Grid Search in Scikit-Learn

Dataset: Coronavirus tweets NLP - Text Classification

The steps to be performed are as follows: Read dataset and perform Text processing for the tweets (Remove Stop words, and special characters and convert the text to lowercase) - 1 Mark Using the train_test_split function of Sklearn, Split the kaggle's train dataset further into train, and test dataset - 1 Mark Use BoW and TF-IDF based feature extraction approaches on the "text" field of the dataset. You can use existing library functions. [2+2 marks] Create model building pipeline and define parameters for GridSearch (You might Refer to the code below) - 2 Mark

text_clf = Pipeline([('vect', CountVectorizer()), ('tfidf', TfidfTransformer()), ('clf', MultinomialNB())])

tuned_parameters = { 'vect_ngram_range': [(1, 1), (1, 2), (2, 2)], 'tfidf_**use_idf': (True, False), 'tfidf**___norm': ('11', '12'), 'clf_alpha': [1, 1e-1, 1e-2] }

5. Perform classification (using GridSearch) - 2 Marks
6. Print the confusion matrix, accuracy, and F1 score on the test dataset - 1 Mark
7. Interpret your results in terms of Business Domain Knowledge. 1 Mark

1.1 Task 1:- Read dataset and perform Text processing for the tweets (Remove Stop words, and special characters and convert the text to lowercase) - 1 Mar

[33]:
```
# Importing required packages
import numpy as np
import pandas as pd
import warnings as war
war.filterwarnings("ignore")
```

[34]:
```
dataSetPath=r"C:\Users\ASUS\jupyterworkspace\Assignment & Mini␣
 ↪Project\Module_05_Text␣
 ↪Mining\Text-Mining-Assignment01-Sentiment-Analysis-Using-Naive-Bayes\Corona_NLP_train.
 ↪csv"
dataSetRead=pd.read_csv(dataSetPath,encoding='ISO-8859-1')
```

[35]:
```
# Displaying first 5 records to confirming data loading
print("**************************************************Displaying below␣
 ↪first 5 records************************************************************")
dataSetRead.head()
```

```
**************************************************Displaying below first 5
records************************************************************
```

[35]:

	UserName	ScreenName	Location	TweetAt	OriginalTweet
0	3799	48751	London	16-03-2020	@MeNyrbie @Phil_Gahan @Chrisitv https://t.co/iFz9FAn2Pa and https://t.co/xX6ghGFzCC and https://t.co/I2NlzdxNo8
1	3800	48752	UK	16-03-2020	advice Talk to your neighbours family to exchange phone numbers create contact list with phone numbers of neighbours schools employer chemist GP set up online shopping accounts if poss adequate supplies of regular meds but not over order
2	3801	48753	Vagabonds	16-03-2020	Coronavirus Australia: Woolworths to give elderly, disabled dedicated shopping hours amid COVID-19 outbreak https://t.co/bInCA9Vp8P
3	3802	48754	NaN	16-03-2020	My food stock is not the only one which is empty…\r\r\n\r\r\nPLEASE, don't panic, THERE WILL BE ENOUGH FOOD FOR EVERYONE if you do not take more than you need. \r\r\nStay calm, stay safe.\r\r\n\r\r\n#COVID19france #COVID_19 #COVID19 #coronavirus #confinement #Confinementotal #ConfinementGeneral https://t.co/zrlG0Z520j
4	3803	48755	NaN	16-03-2020	

```
4  Me, ready to go at supermarket during the #COVID19 outbreak.\r\r\n\r\r\nNot
because I'm paranoid, but because my food stock is litteraly empty. The
#coronavirus is a serious thing, but please, don't panic. It causes
shortage…\r\r\n\r\r\n#CoronavirusFrance #restezchezvous #StayAtHome
#confinement https://t.co/usmuaLq72n

            Sentiment
0             Neutral
1            Positive
2            Positive
3            Positive
4  Extremely Negative
```

[36]:
```
# Displaying dimension of dataSet
print("Dimention of Dataset:- {}".format(dataSetRead.shape[0:2]))
print("Total number of rows in Dataset:- {}".format(dataSetRead.shape[0]))
print("Total number of columns in Dataset:- {}".format(dataSetRead.shape[1]))
```

```
Dimention of Dataset:- (41157, 6)
Total number of rows in Dataset:- 41157
Total number of columns in Dataset:- 6
```

[37]:
```
# Selecting relevent features from dataSet
dataSetRead=dataSetRead[['OriginalTweet','Sentiment']]
```

1.1.1 Removing Stopwords

[38]:
```
# Importing required packages
from nltk.corpus import stopwords # nltk.corpus.stopwords: Provides a␣
 ↪collection of common stopwords for multiple languages.
from nltk.tokenize import word_tokenize # nltk.tokenize.word_tokenize: A␣
 ↪tokenizer that splits text into individual words
import nltk # import nltk: The Natural Language Toolkit is a library used for␣
 ↪natural language processing tasks.

# Downloading the stopwords and punkt tokenizer if not already downloaded
nltk.download('stopwords') # nltk.download('stopwords'): Ensures the required␣
 ↪stopword dataset is downloaded locally
nltk.download('punkt') # nltk.download('punkt'): Downloads the punkt tokenizer␣
 ↪model, which is needed for tokenizing text into words or sentences

# Getting the list of English stopwords
stop_words = set(stopwords.words('english')) # stopwords.words('english'):␣
 ↪Retrieves a predefined list of English stopwords

# Functioning to remove stopwords
```

```
def remove_stopwords(text): # def remove_stopwords(text):: Defines a function␣
 ↪to clean text by removing stopwords
    if not isinstance(text, str): # if not isinstance(text, str):: Checks if␣
 ↪the input is a string; if not, it returns the input unchanged
        return text
    words = word_tokenize(text)
    filtered_words = [word for word in words if word.lower() not in stop_words]
    return ' '.join(filtered_words)

# Applying the function to the 'OriginalTweet' column
dataSetRead['text_cleaned_OriginalTweet'] = dataSetRead['OriginalTweet'].
 ↪apply(remove_stopwords)
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\ASUS\AppData\Roaming\nltk_data…
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\ASUS\AppData\Roaming\nltk_data…
[nltk_data]   Package punkt is already up-to-date!
```

[39]:
```
pd.set_option('display.max_colwidth', None)  # Show full cell content for long␣
 ↪text
dataSetRead['text_cleaned_OriginalTweet']
```

[39]:
```
0
@ MeNyrbie @ Phil_Gahan @ Chrisitv https : //t.co/iFz9FAn2Pa https :
//t.co/xX6ghGFzCC https : //t.co/I2NlzdxNo8
1                                                                advice Talk
neighbours family exchange phone numbers create contact list phone numbers
neighbours schools employer chemist GP set online shopping accounts poss
adequate supplies regular meds order
2
Coronavirus Australia : Woolworths give elderly , disabled dedicated shopping
hours amid COVID-19 outbreak https : //t.co/bInCA9Vp8P
3                        food stock one empty … PLEASE , n't panic , ENOUGH
FOOD EVERYONE take need . Stay calm , stay safe . # COVID19france # COVID_19 #
COVID19 # coronavirus # confinement # Confinementotal # ConfinementGeneral https
: //t.co/zrlG0Z520j
4         , ready go supermarket # COVID19 outbreak . 'm paranoid , food stock
litteraly empty . # coronavirus serious thing , please , n't panic . causes
shortage … # CoronavirusFrance # restezchezvous # StayAtHome # confinement
https : //t.co/usmuaLq72n
                                                      …
41152
Airline pilots offering stock supermarket shelves # NZ lockdown # COVID-19 https
: //t.co/cz89uAOHNp
41153
```

```
Response complaint provided citing COVID-19 related delays . Yet prompt
rejecting policy consumer TAT . Way go ?
41154
know itÂ s getting tough @ KameronWilds rationing toilet paper # coronavirus #
toiletpaper @ kroger martinsville , help us ! !
41155
wrong smell hand sanitizer starting turn ? # coronavirus # COVID19 # coronavirus
41156                                                    @ TartiiCat Well
new/used Rift going $ 700.00 Amazon rn although normal market price usually $
400.00 . Prices really crazy right vr headsets since HL Alex announced 's worse
COVID-19 . whethe
Name: text_cleaned_OriginalTweet, Length: 41157, dtype: object
```

1.1.2 Removing the special characters and convert the text to lower case.

[40]:
```
# Importing required package
import re

# Functioning to preprocess text
def preprocess_text(text):
# Removing punctuation and special characters using regex
    text = re.sub(r'[^a-zA-Z\s]', '', text)  # Keeping only words and spaces
# Converting text to lowercase
    text = text.lower()
    return text

# Applying the preprocessing function to the 'OriginalTweet' column
dataSetRead['text_cleaned_OriginalTweet'] =␣
 ↪dataSetRead['text_cleaned_OriginalTweet'].apply(preprocess_text)

# Displaying the 10 five rocords of updated dataFrame
print("**********************************************Displaying below the␣
 ↪first 10 rocords of updated␣
 ↪dataframe**********************************************")
dataSetRead.head(20)
```

```
**********************************************Displaying below the first 10
rocords of updated dataframe**********************************************
```

[40]:
```
   OriginalTweet  \
0
@MeNyrbie @Phil_Gahan @Chrisitv https://t.co/iFz9FAn2Pa and
https://t.co/xX6ghGFzCC and https://t.co/I2NlzdxNo8
1
advice Talk to your neighbours family to exchange phone numbers create contact
list with phone numbers of neighbours schools employer chemist GP set up online
shopping accounts if poss adequate supplies of regular meds but not over order
```

2
Coronavirus Australia: Woolworths to give elderly, disabled dedicated shopping hours amid COVID-19 outbreak https://t.co/bInCA9Vp8P
3 My food stock is not the only one which is empty…\r\r\n\r\r\nPLEASE, don't panic, THERE WILL BE ENOUGH FOOD FOR EVERYONE if you do not take more than you need. \r\r\nStay calm, stay safe.\r\r\n\r\r\n#COVID19france #COVID_19 #COVID19 #coronavirus #confinement #Confinementotal #ConfinementGeneral https://t.co/zrlG0Z520j
4 Me, ready to go at supermarket during the #COVID19 outbreak.\r\r\n\r\r\nNot because I'm paranoid, but because my food stock is litteraly empty. The #coronavirus is a serious thing, but please, don't panic. It causes shortage…\r\r\n\r\r\n#CoronavirusFrance #restezchezvous #StayAtHome #confinement https://t.co/usmuaLq72n
5 As news of the regionÂ s first confirmed COVID-19 case came out of Sullivan County last week, people flocked to area stores to purchase cleaning supplies, hand sanitizer, food, toilet paper and other goods, @Tim_Dodson reports https://t.co/cfXch7a2lU
6
Cashier at grocery store was sharing his insights on #Covid_19 To prove his credibility he commented "I'm in Civics class so I know what I'm talking about". https://t.co/ieFDNeHgDO
7
Was at the supermarket today. Didn't buy toilet paper. #Rebel\r\r\n\r\r\n#toiletpapercrisis #covid_19 https://t.co/eVXkQLIdAZ
8 Due to COVID-19 our retail store and classroom in Atlanta will not be open for walk-in business or classes for the next two weeks, beginning Monday, March 16. We will continue to process online and phone orders as normal! Thank you for your understanding! https://t.co/kw91zJ5O5i
9 For corona prevention,we should stop to buy things with the cash and should use online payment methods because corona can spread through the notes. Also we should prefer online shopping from our home. It's time to fight against COVID 19?. #govindia #IndiaFightsCorona
10 All month there hasn't been crowding in the supermarkets or restaurants, however reducing all the hours and closing the malls means everyone is now using the same entrance and dependent on a single supermarket. #manila #lockdown #covid2019 #Philippines https://t.co/HxWs9LAnF9
11
Due to the Covid-19 situation, we have increased demand for all food products. \r\r\n\r\r\nThe wait time may be longer for all online orders, particularly beef share and freezer packs. \r\r\n\r\r\nWe thank you for your patience during this time.
12 #horningsea is a caring community. LetÂ s ALL look after the less capable in our village and ensure they

stay healthy. Bringing shopping to their doors, help with online shopping and self isolation if you have symptoms or been exposed to somebody who has. https://t.co/lsGrXXhjhh
13
Me: I don't need to stock up on food, I'll just have Amazon deliver whatever I need #CoronaVirus\r\r\n\r\r\nAmazon: https://t.co/8YWaKFjExC
14
ADARA Releases COVID-19 Resource Center for Travel Brands: Insights Help Travel Brands Stay Up-To-Date on Consumer Travel Behavior Trends https://t.co/PnA797jDKV https://t.co/dQox6uSihz
15 Lines at the grocery store have been unpredictable, but is eating out a safe alternative? \r\r\n\r\r\nFind out more about whether you should be avoiding restaurants right now: https://t.co/9idZSis5oQ\r\r\n\r\r\n#coronavirus #covid19 https://t.co/ZHbh898lf6
16
????? ????? ????? ????? ??\r\r\n?????? ????? ??????? ????????? ?\r\r\n#????_???? ????? ???? ?????? ? #????????? ?????? ?? 13 ???? ?? ?? ???? ?????? ?? ??\r\r\n#???_???????? ????? ??? ???? ? https://t.co/51bL8P6vZh
17
@eyeonthearctic 16MAR20 Russia consumer surveillance watchdog reported case in high Arctic where a man who traveled to Iran has COVID-19 and 101 are "observed"\r\r\nhttps://t.co/4WnrrK9oKC https://t.co/ld05k5Eyns
18 Amazon Glitch Stymies Whole Foods, Fresh Grocery Deliveries\r\r\nÂ As COVID-19 has spread, weÂ ve seen a significant increase in people shopping online for groceries,Â a spokeswoman said in a statement. Â Today this resulted in a systems impact affecting our …\r\r\n https://t.co/TbzZ2MC3b3
19
For those who aren't struggling, please consider donating to a food bank or a nonprofit. The demand for these services will increase as COVID-19 impacts jobs, and people's way of life.

	Sentiment
0	Neutral
1	Positive
2	Positive
3	Positive
4	Extremely Negative
5	Positive
6	Positive
7	Neutral
8	Positive
9	Negative
10	Neutral
11	Extremely Positive
12	Extremely Positive

```
13            Positive
14            Positive
15            Positive
16             Neutral
17             Neutral
18  Extremely Positive
19            Positive

                                                text_cleaned_OriginalTweet
0
menyrbie  philgahan  chrisitv https  tcoifzfanpa https  tcoxxghgfzcc https
tcoinlzdxno
1                                      advice talk neighbours family exchange
phone numbers create contact list phone numbers neighbours schools employer
chemist gp set online shopping accounts poss adequate supplies regular meds
order
2
coronavirus australia  woolworths give elderly  disabled dedicated shopping
hours amid covid outbreak https  tcobincavpp
3                                  food stock one empty  please  nt panic  enough
food everyone take need  stay calm  stay safe   covidfrance  covid  covid
coronavirus  confinement  confinementotal  confinementgeneral https  tcozrlgzj
4             ready go supermarket  covid outbreak  m paranoid  food stock
litteraly empty   coronavirus serious thing  please  nt panic  causes shortage
coronavirusfrance  restezchezvous  stayathome  confinement https  tcousmualqn
5                           news regions first confirmed covid case came sullivan
county last week  people flocked area stores purchase cleaning supplies  hand
sanitizer  food  toilet paper goods   timdodson reports https  tcocfxchalu
6
cashier grocery store sharing insights  covid prove credibility commented  m
civics class know m talking   https  tcoiefdnehgdo
7
supermarket today  nt buy toilet paper   rebel  toiletpapercrisis  covid https
tcoevxkqlidaz
8                                        due covid retail store classroom atlanta
open walkin business classes next two weeks  beginning monday  march   continue
process online phone orders normal  thank understanding  https  tcokwzjoi
9                                                          corona prevention
stop buy things cash use online payment methods corona spread notes  also prefer
online shopping home  s time fight covid     govindia  indiafightscorona
10                                    month nt crowding supermarkets restaurants
however reducing hours closing malls means everyone using entrance dependent
single supermarket   manila  lockdown  covid  philippines https  tcohxwslanf
11
due covid situation  increased demand food products  wait time may longer online
orders  particularly beef share freezer packs  thank patience time
12                                           horningsea caring community  lets
```

```
look less capable village ensure stay healthy  bringing shopping doors  help
online shopping self isolation symptoms exposed somebody  https  tcolsgrxxhjhh
13
nt need stock food  ll amazon deliver whatever need  coronavirus amazon  https
tcoywakfjexc
14                                                                    adara
releases covid resource center travel brands  insights help travel brands stay
uptodate consumer travel behavior trends https  tcopnajdkv https  tcodqoxusihz
15
lines grocery store unpredictable  eating safe alternative  find whether
avoiding restaurants right  https  tcoidzsisoq  coronavirus  covid https
tcozhbhlf
16
https  tcoblpvzh
17
eyeonthearctic mar russia consumer surveillance watchdog reported case high
arctic man traveled iran covid  observed  https  tcownrrkokc https  tcoldkeyns
18  amazon glitch stymies whole foods  fresh grocery deliveries as covid spread
weve seen significant increase people shopping online groceries  spokeswoman
said statement  today resulted systems impact affecting  https  tcotbzzmcb
19
nt struggling  please consider donating food bank nonprofit  demand services
increase covid impacts jobs  people s way life
```

1.2 Task 2:- Using the train_test_split function of Sklearn, Split the kaggle's train dataset further into train, and test dataset - 1 Mark

[41]:
```
# Importing required package
from sklearn.model_selection import train_test_split
# Defining the feature (X) and the target variable (y)
X=dataSetRead['OriginalTweet']
y=dataSetRead['Sentiment']
# Split the dataset into training and testing subsets
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.
 ↪3,random_state=42)
# Displaying the shapes of the resulting datasets
print("X_train shape:", X_train.shape[0])
print("X_test shape:", X_test.shape[0])
print("y_train shape:", y_train.shape[0])
print("y_test shape:", y_test.shape[0])
```

```
X_train shape: 28809
X_test shape: 12348
y_train shape: 28809
y_test shape: 12348
```

1.2.1 Task 3:- Use BoW and TF-IDF based feature extraction approaches on the "text" field of the dataset. You can use existing library functions. [2+2 marks]

[42]:
```
# Importing required package
from sklearn.feature_extraction.text import CountVectorizer

# Step 1: Initialize CountVectorizer
bow_vectorizer = CountVectorizer()

# Step 2: Fit the vectorizer on training data to learn the vocabulary
bow_vectorizer.fit(X_train)

# Step 3: Transform both the training and test data
X_train_bow = bow_vectorizer.transform(X_train)
X_test_bow = bow_vectorizer.transform(X_test)

# Step 4: Print the shape of the transformed matrices
print(f"BoW - Training Data Shape: {X_train_bow.shape}")
print(f"BoW - Testing Data Shape: {X_test_bow.shape}")
```

```
BoW - Training Data Shape: (28809, 62185)
BoW - Testing Data Shape: (12348, 62185)
```

[43]:
```
sparsity = (X_train_bow.nnz / (X_train_bow.shape[0] * X_train_bow.shape[1])) *␣
↪100
print(f"Sparsity: {sparsity:.2f}%")
```

```
Sparsity: 0.04%
```

[44]:
```
# Importing required package

from sklearn.feature_extraction.text import TfidfVectorizer

# Step 1: Initialize TfidfVectorizer
tfidf_vectorizer = TfidfVectorizer()

# Step 2: Fit the vectorizer on the training data to learn the vocabulary and␣
↪IDF values
tfidf_vectorizer.fit(X_train)

# Step 3: Transform the training and test data into TF-IDF matrices
X_train_tfidf = tfidf_vectorizer.transform(X_train)
X_test_tfidf = tfidf_vectorizer.transform(X_test)

# Step 4: Print the shape of the transformed matrices
print(f"TF-IDF - Training Data Shape: {X_train_tfidf.shape}")
print(f"TF-IDF - Testing Data Shape: {X_test_tfidf.shape}")
```

```
TF-IDF - Training Data Shape: (28809, 62185)
TF-IDF - Testing Data Shape: (12348, 62185)
```

[45]:
```
sparsity = (X_train_tfidf.nnz / (X_train_tfidf.shape[0] * X_train_tfidf.
 ↪shape[1])) * 100
print(f"Sparsity: {sparsity:.2f}%")
```

```
Sparsity: 0.04%
```

1.3 Task 4:- Create model building pipeline and define parameters for GridSearch (You might Refer to the code below) - 2 Mark

text_clf = Pipeline([('vect', CountVectorizer()), ('tfidf', TfidfTransformer()), ('clf', MultinomialNB())])

tuned_parameters = { 'vect__ngram_range': [(1, 1), (1, 2), (2, 2)], 'tfidf__**use_idf': (True, False), 'tfidf**__norm': ('11', '12'), 'clf__alpha': [1, 1e-1, 1e-2] }

[46]:
```
# Importing required packages
from sklearn.pipeline import Pipeline
from sklearn.feature_extraction.text import CountVectorizer, TfidfTransformer
from sklearn.naive_bayes import MultinomialNB

# Define the pipeline
text_clf = Pipeline([
    ('vect', CountVectorizer()),  # Convert text to a matrix of token counts
    ('tfidf', TfidfTransformer()),  # Transform counts to TF-IDF representation
    ('clf', MultinomialNB())  # Apply Multinomial Naive Bayes for classification
])

# Define the parameter grid for GridSearchCV
tuned_parameters = {
    'vect__ngram_range': [(1, 1), (1, 2), (2, 2)],  # N-gram range for␣
 ↪tokenization
    'tfidf__use_idf': [True, False],  # Whether to use IDF weighting
    'tfidf__norm': ['l1', 'l2'],  # Normalization options
    'clf__alpha': [1, 0.1, 0.01]  # Smoothing parameter for MultinomialNB
}

# Note:
# - Double underscores (`__`) are used to access the parameters of pipeline␣
 ↪components.
# - Corrected parameter names to match proper syntax (`vect__ngram_range`,␣
 ↪`tfidf__use_idf`, etc.).
```

[47]:
```
# Importing required package
from sklearn.model_selection import GridSearchCV
```

```
# Set up GridSearchCV
grid_search = GridSearchCV(text_clf, tuned_parameters, scoring='accuracy',␣
 ↪cv=5, verbose=1)

# Fit the model on the training data
grid_search.fit(X_train,y_train)

# Display the best parameters and best score
print("Best Parameters:", grid_search.best_params_)
print("Best Cross-Validation Score:", grid_search.best_score_)
```

```
Fitting 5 folds for each of 36 candidates, totalling 180 fits
Best Parameters: {'clf__alpha': 0.01, 'tfidf__norm': 'l2', 'tfidf__use_idf':
False, 'vect__ngram_range': (1, 2)}
Best Cross-Validation Score: 0.44579823480017194
```

1.4 Task 5:- Perform classification (using GridSearch) - 2 Marks

[48]:
```
# Importing required packages
from sklearn.metrics import accuracy_score, confusion_matrix, f1_score
from sklearn.metrics import classification_report
# Get the best model from GridSearchCV
bestModel = grid_search.best_estimator_
# Predict the sentiment on the test data
y_pred = bestModel.predict(X_test)
print("Classification Report on Test Set:")
print(classification_report(y_test, y_pred))
```

```
Classification Report on Test Set:
                    precision    recall  f1-score   support

Extremely Negative       0.55      0.36      0.44      1572
Extremely Positive       0.55      0.40      0.46      1989
          Negative       0.41      0.48      0.44      3005
           Neutral       0.65      0.39      0.48      2292
          Positive       0.41      0.58      0.48      3490

          accuracy                           0.46     12348
         macro avg       0.51      0.44      0.46     12348
      weighted avg       0.49      0.46      0.46     12348
```

1.5 Task 6 :- Print the confusion matrix, accuracy, and F1 score on the test dataset - 1 Mark

[49]:
```
# Importing required package
from sklearn.metrics import confusion_matrix, accuracy_score, f1_score

# Step 1: Predict on the test set
y_pred = bestModel.predict(X_test)

# Step 2: Compute the confusion matrix
conf_matrix = confusion_matrix(y_test, y_pred)
print("Confusion Matrix:")
print(conf_matrix)

# Step 3: Compute accuracy
accuracy = accuracy_score(y_test, y_pred)
print(f"Accuracy: {accuracy:.4f}")

# Step 4: Compute F1 score (you can also specify 'weighted' if you have␣
 ↪imbalanced classes)
f1 = f1_score(y_test, y_pred, average='weighted')
print(f"F1 Score (Weighted): {f1:.4f}")
```

```
Confusion Matrix:
[[ 566   21  709   36  240]
 [  20  791  169   49  960]
 [ 310  109 1437  176  973]
 [  56   85  486  887  778]
 [  71  445  726  220 2028]]
Accuracy: 0.4623
F1 Score (Weighted): 0.4620
```

1.6 Task 7:- Interpret your results in terms of Business Domain Knowledge. 1 Mark

In the confusion matrix, the rows represent the true labels, and the columns represent the predicted labels.

Key Observations: Class 0 (True vs. Predicted):

There are 566 instances correctly classified as Class 0 (True Positives). 709 instances were incorrectly classified as Class 2, 240 instances as Class 4, and only 36 instances as Class 3. A significant portion of the instances from Class 0 is being misclassified into Class 2. Class 1:

791 instances correctly classified as Class 1 (True Positives). Misclassification into other classes: 169 into Class 2, 960 into Class 4. This suggests that the model struggles to distinguish Class 1 from Class 2 and Class 4. Class 2:

1437 instances are correctly identified as Class 2 (True Positives). A fair number of misclassifications into Class 0 (310) and Class 1 (109), but still a relatively high number of true positives. Class 3:

887 instances are correctly classified as Class 3 (True Positives). 486 instances misclassified into Class 2 and 778 into Class 4, indicating significant confusion with these classes. Class 4:

2028 instances correctly classified as Class 4 (True Positives). Misclassifications into Class 1 (445) and Class 2 (726) are notable, with 220 instances incorrectly classified as Class 3. Accuracy: 0.4623 Interpretation: The model has an accuracy of 46.23%, meaning the model correctly predicted the class for about 46.23% of the instances. This is a relatively low accuracy, indicating that the model might be struggling to distinguish between certain classes, or the dataset might be highly imbalanced. Business Insight: If this is a classification task where accurate predictions are critical (e.g., fraud detection, customer segmentation), this low accuracy may not be sufficient for reliable decision-making. Further improvements in feature engineering, model tuning, or data preprocessing are needed. F1 Score (Weighted): 0.4620 Interpretation: The F1 score (weighted) is 0.4620, which is close to the accuracy. The weighted F1 score accounts for class imbalance by giving higher weight to the classes with more instances. This score suggests that the model performs poorly across the classes, as the F1 score considers both precision and recall (balancing false positives and false negatives). Business Insight: This F1 score suggests that the model is not effective at identifying positive cases in many classes, and may have a high rate of both false positives and false negatives. Improving this score should be a priority if the model is to be used in business applications where missing a true positive or flagging too many false positives can have significant costs. Potential Business Implications and Actions: Class Imbalance:

If the dataset is imbalanced (i.e., some classes have many more samples than others), this can cause the model to favor the majority class. Consider using class weighting or techniques like SMOTE (Synthetic Minority Over-sampling Technique) to balance the dataset. Model Tuning:

The model might not be tuned optimally. Consider running a GridSearchCV or RandomizedSearchCV to optimize hyperparameters and improve performance. Feature Engineering:

Review the features being used to train the model. It might be necessary to extract more informative features, handle missing data, or remove irrelevant ones. Additional Metrics:

Depending on the business use case, consider evaluating precision, recall, or ROC AUC scores for each class individually. This would provide more granular insights into the model's performance across different classes, especially in cases where false negatives or false positives are more costly. Model Choice:

You may want to try different classifiers (e.g., Random Forest, XGBoost, SVM) to see if other models perform better than Naive Bayes in distinguishing between the classes.

Next Steps: Data Cleaning and Preprocessing: Investigate potential issues with data quality, such as incorrect labels or noisy data. Improve Feature Selection: Consider using more domain-specific features or advanced techniques like TF-IDF or word embeddings (if it's text-based data). Hyperparameter Tuning: Fine-tune the model to improve its performance, using techniques like cross-validation to get more stable results.