# Data Collection and Preprocessing Phase

| | |
|---|---|
| Date | June 2024 |
| Team ID | 740096 |
| Project Title | The Language Of You tube: A Text Classification Approach To Video Descriptions |
| Maximum Marks | 6 Marks |

**Preparation Template**

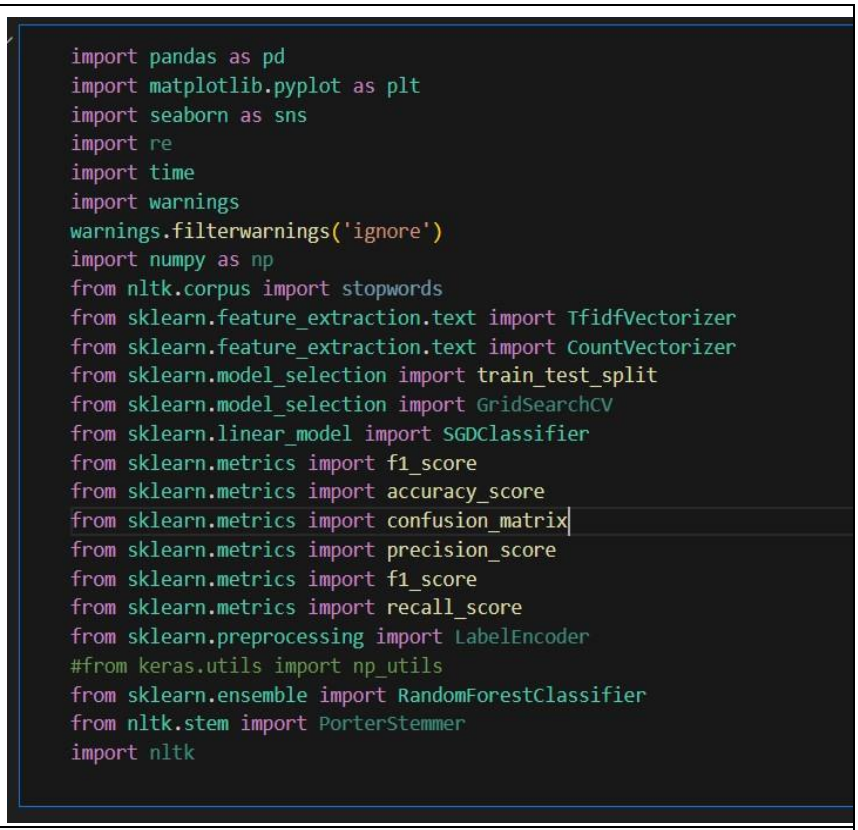
The images will be pre processed by resizing, normalizing, augmenting, denoising, adjusting contrast, detecting edges, converting color space, cropping, batch normalizing, and whitening data. These steps will enhance data quality, promote model generalization, and improve convergence during neural network training, ensuring robust and efficient performance across various computer vision tasks.

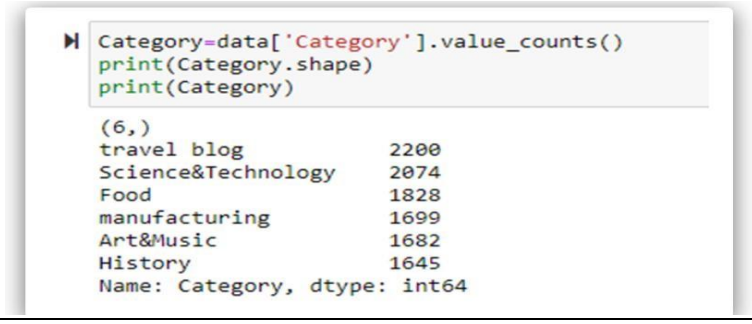
| Section | Description |
|---|---|
| Data Overview | There are many popular open sources for collecting the data. Eg: kaggle.com, UCI repository, etc. In this project we have used .csv data. |
| Data Preparation | These are the general steps of pre-processing the data before using it for machine learning |

| | |
|---|---|
| Handling missing values | We use Handling missing values For checking the null values |
| Handling categorical data | As we can see our dataset has categorical data we must convert the categorical data to integer encoding or binary encoding |
| Handling Outliers in Data | With the help of boxplot, outliers are visualized. And here we are going to find upper bound and lower bound of numerical features with some mathematical formula. |

# Preparation

| | |
|---|---|
| Collect the dataset | Please refer to the link given below to download the dataset.<br>Youtube Videos Dataset (~3400 videos) (kaggle.com) |

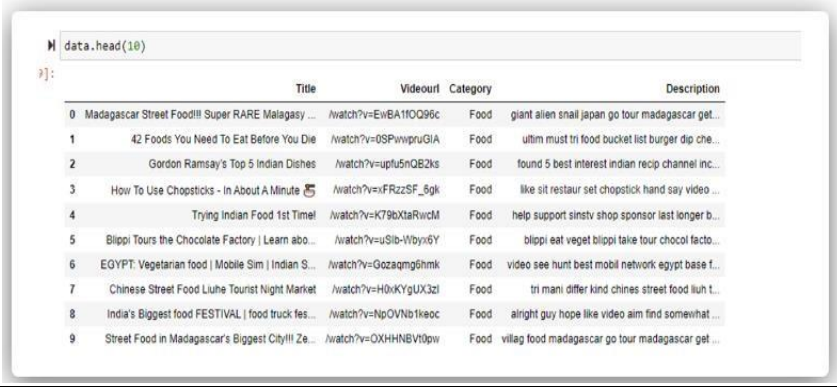
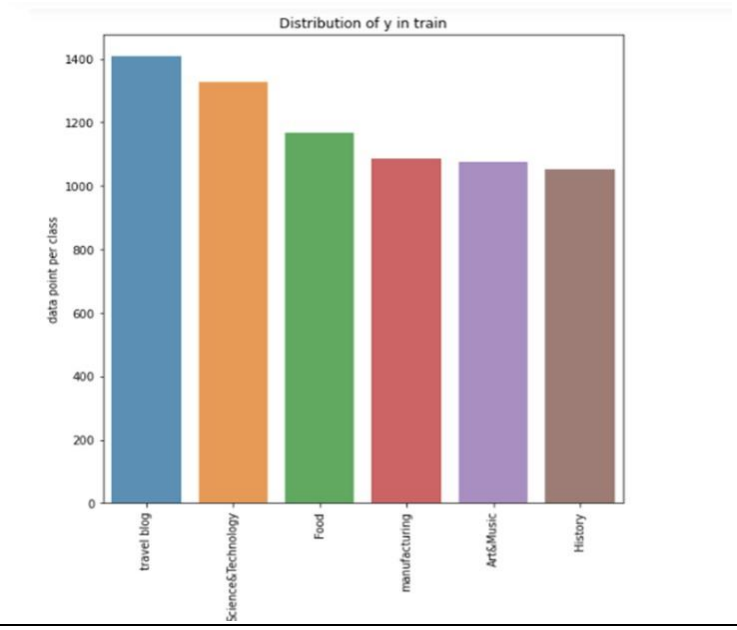
| | |
|---|---|
| Importing the libraries |  |
| Loading Data | We use the code<br>Data =pd.read_csv('YoutubeDataSet.csv')<br>For reading the dataset |

| | |
|---|---|
| Handling missing values |  |

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import re
import time
import warnings
warnings.filterwarnings('ignore')
import numpy as np
from nltk.corpus import stopwords
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.model_selection import train_test_split
from sklearn.model_selection import GridSearchCV
from sklearn.linear_model import SGDClassifier
from sklearn.metrics import f1_score
from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix
from sklearn.metrics import precision_score
from sklearn.metrics import f1_score
from sklearn.metrics import recall_score
from sklearn.preprocessing import LabelEncoder
#from keras.utils import np_utils
from sklearn.ensemble import RandomForestClassifier
from nltk.stem import PorterStemmer
import nltk
```

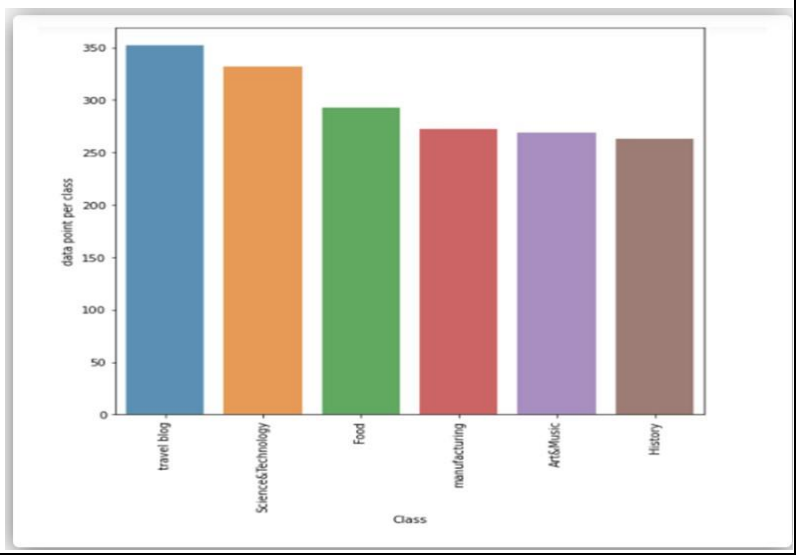
```
Category=data['Category'].value_counts()
print(Category.shape)
print(Category)
```

```
(6,)
travel blog            2200
Science&Technology     2074
Food                   1828
manufacturing          1699
Art&Music              1682
History                1645
Name: Category, dtype: int64
```

| | |
|---|---|
| **Preprocessing of Description** |  |
| **Visualizing the distribution of target variable in the test dataset** |  |

```
data.head(10)
```

| | Title | Videourl | Category | Description |
|---|---|---|---|---|
| 0 | Madagascar Street Food!!! Super RARE Malagasy ... | /watch?v=EwBA1fOQ96c | Food | giant alien snail japan go tour madagascar get... |
| 1 | 42 Foods You Need To Eat Before You Die | /watch?v=0SPwwpruGIA | Food | ultim must tri food bucket list burger dip che... |
| 2 | Gordon Ramsay's Top 5 Indian Dishes | /watch?v=upfu5nQB2ks | Food | found 5 best interest indian recip channel inc... |
| 3 | How To Use Chopsticks - In About A Minute | /watch?v=xFRzzSF_6gk | Food | like sit restaur set chopstick hand say video ... |
| 4 | Trying Indian Food 1st Time! | /watch?v=K79bXtaRwcM | Food | help support sinstv shop sponsor last longer b... |
| 5 | Blippi Tours the Chocolate Factory \| Learn abo... | /watch?v=uSlb-WbyxY | Food | blippi eat veget blippi take tour chocol facto... |
| 6 | EGYPT: Vegetarian food \| Mobile Sim \| Indian S... | /watch?v=Gozaqmg6hmk | Food | video see hunt best mobil network egypt base f... |
| 7 | Chinese Street Food Liuhe Tourist Night Market | /watch?v=H0xKYgUX3zI | Food | tri mani differ kind chines street food liuh t... |
| 8 | India's Biggest food FESTIVAL \| food truck fes... | /watch?v=NpOVNb1keoc | Food | alright guy hope like video aim find somewhat ... |
| 9 | Street Food in Madagascar's Biggest City!!! Ze... | /watch?v=OXHHNBVt0pw | Food | villag food madagascar go tour madagascar get ... |


data point per class
0
50
100
150
200
250
300
350
travel blog
Science&Technology
Food
manufacturing
Art&Music
History
Class