MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR

(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)

Minor Project Report

on

News Classification Using Machine learning

A project report submitted in partial fulfillment of the requirement for the degree of

BACHELOR OF TECHNOLOGY

In

Computer Science & Engineering

Submitted by:
Utkarsh Saxena
0901CS201131

FacultyMentor:
Dr. Ranjeet Kumar Singh
Assistant Professor,CSE

Department Of Computer Science & Engineering
Madhav Institute Of Technology & Science Gwalior –
474005,(Mp) Est. 1957
Jan-June 2022

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR

(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)

CERTIFICATE

This is certified that Utkarsh Saxena (0901CS201131) has submitted the project report titled News Classification using ML under the mentorship of Dr. Ranjeet Kumar Singh, in partial fulfillment of the requirement for the award of degree of Bachelor of Technology in Computer Science &Engineering from Madhav Institute of Technology and Science ,Gwalior.

R.K.Singh

Dr. Ranjeet Kumar Singh

Assistant Professor
CSE Department

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR

(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal

DECLARATION

I here by declare that the work being presented in this project report , for the partial fulfilment of requirement for the award of the degree of Bachelor of Technology in Computer Science at Madhav Institute of Technology & Science, Gwalior is an authenticated and original record of my work under the mentorship of Dr. Ranjeet Kumar Singh.

I declare that I have not submitted the matter embodied in this report for the award of any degree or diploma anywhere else.

UtkarshSaxena
0901CS201131

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR

(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)

ACKNOWLEDGEMENT

The full semester project has proved to be pivotal to my career. I am thankful to my institute, Madhav Institute of Technology and Science to allow me to continue my disciplinary/interdisciplinary project as a curriculum requirement, under the provisions of the Flexible Curriculum Scheme (based on the AICTE Model Curriculum 2018), approved by the Academic Council of the institute. I extend my gratitude to the Director of the institute, Dr. R. K. Pandit and Dean Academics, Dr. Manjaree Pandit for this.

I would sincerely like to thank my department,Department of ComputerScience, for allowing me to explore this project. I humbly thank Dr.Manish Dixit, Professor and Head, Department of Computer Science & Engineering, for his continued support during the course of this engagement,which eased the process and formalities involved.

I am sincerely thankful to my faculty mentors. I am grateful to the guidance Dr. Ranjeet Kumar Singh for their continued support and guidance throughout the project . I am also very thankful to the faculty and staff of the department.

Usaxena

UtkarshSaxena
0901CS201131

Abstract

This machine learning model is based on the principle of supervised learning which includes classification (grouping) of data in the predefined sets. This project can be used to categories new upcoming news in the following categories hence reducing time and labor and making machine capable enough to make decisions.

It can be used widely by news channels and in research purposes

The Multinomial Naive Bayes algorithm is a Bayesian learning approach popular in Natural Language Processing (NLP). The program guesses the tag of a text, such as an email or a newspaper story, using the Bayes theorem. It calculates each tag's likelihood for a given sample and outputs the tag with the greatest chance

In this project I have used Python programming language for machine learning. This project can be used widely by TV reporters and also for research purposes in the field of machine learning for segregation of data. It is under progress and is very easy to implement.

It uses multinomial naïve bayes algorithm for NLP.

INDEX

CHAPTER 1 :- INRODUCTION

You must have seen the news divided into categories when you go to a news website. Some of the popular categories that you'll see on almost any news website are tech, entertainment , and sports . If you want to know how to classify news categories using machine learning, this article is for you. In this Introduction , I will walk you through the task of news classification with machine learning using Python.

Every news website classifies the news article before publishing it so that every time visitors visit their website can easily click on the type of news that interests them. For example, I like to read the latest technology updates, so every time I visit a news website, I click on the technology section. But you may or may not like to read about technology ,you may be interested in politics ,business, entertainment, or may be sports.

Currently, the news articles are classified by hand by the content managers of news websites. But to save time, they can also implement a machine learning model on their websites that read the news headline or the content of the news and classifies the category of the news. In the section below, I will take you through how you can train a machine learning model for the task of news classification using the Python programming language.

CHAPTER 2 :- OBJECTIVE

Our aim is to make a model which can classify the up coming news based on previously news

HARDWARE

1. Harddisk–100GB.
2. RAM–1GB
3. OS–Windows ,Linux

LANGUAGE&SOFTWARE

1. Python–Pandas ,NumPy,matplotlib,seaborn,sklearn.
2. Dataset–News classification
3. IDE–Jupyter,Colab, anaconda.

CHAPTER 3 :- FLOWCHART

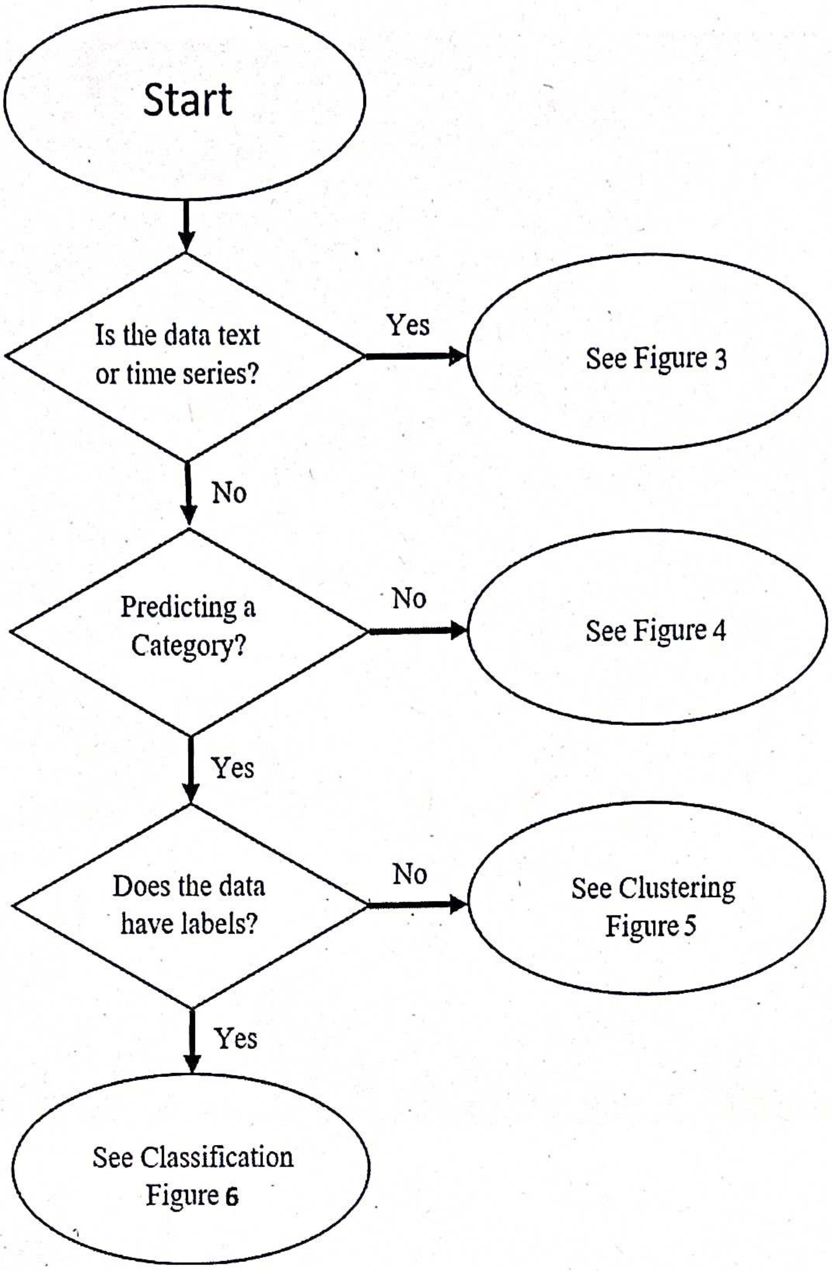
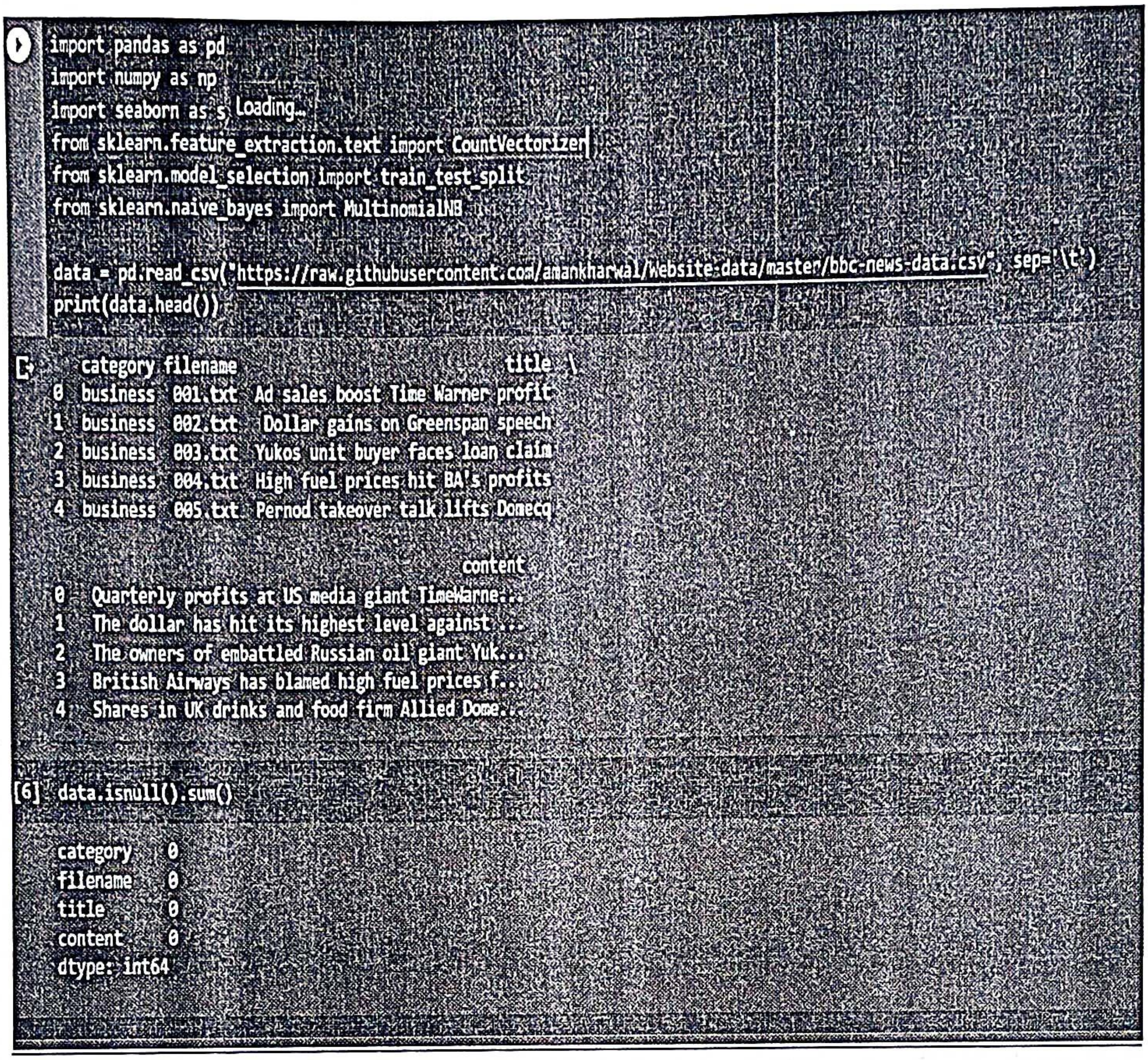


Fig 3.1 Working structure of the project

STRUCTURE OF THE PROGRAM

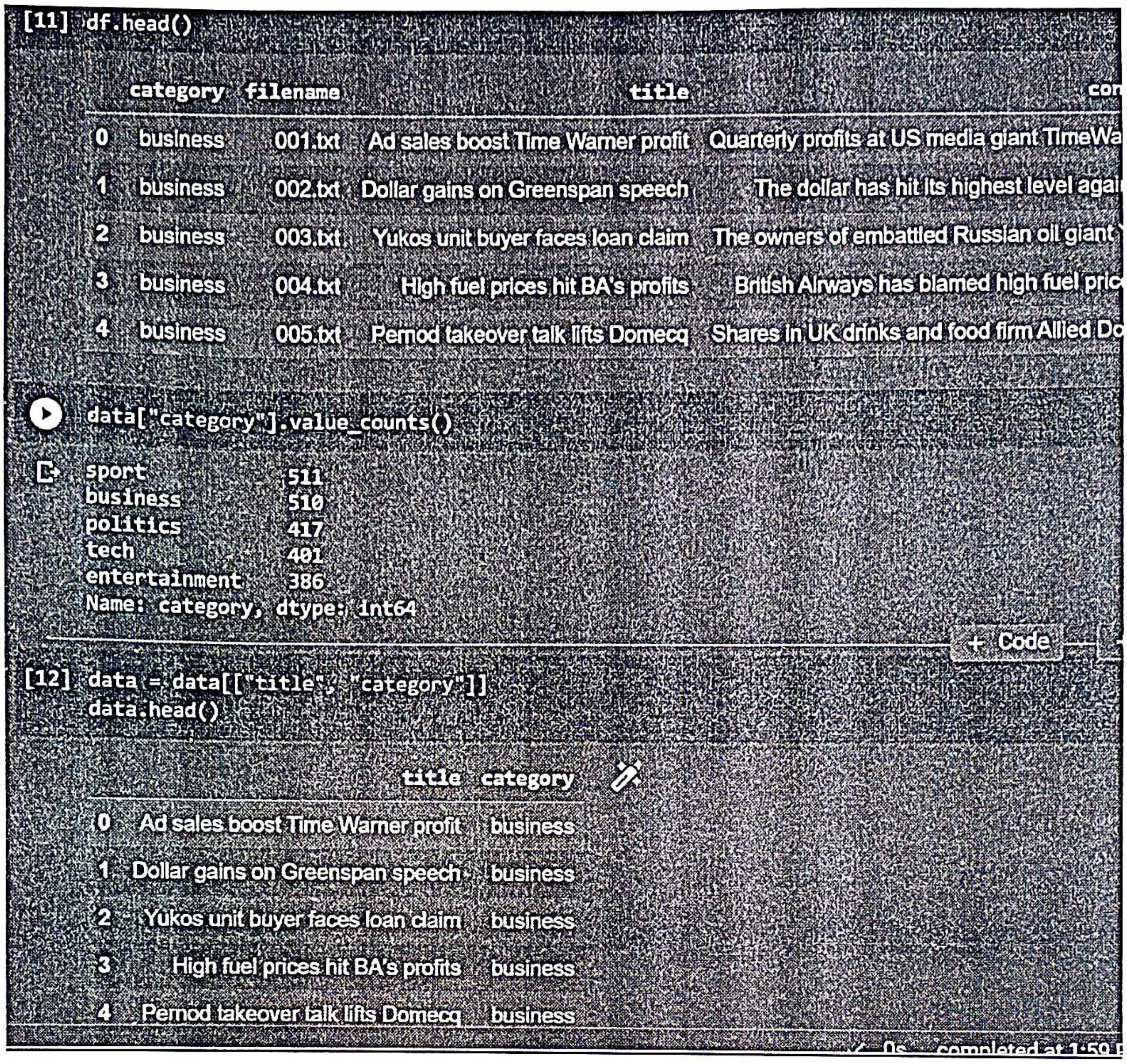


```
import pandas as pd
import numpy as np
import seaborn as s Loading...
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import MultinomialNB

data = pd.read_csv("https://raw.githubusercontent.com/amankharwal/Website-data/master/bbc-news-data.csv", sep='\t')
print(data.head())
```

```
  category filename                            title  \
0 business  001.txt  Ad sales boost Time Warner profit
1 business  002.txt   Dollar gains on Greenspan speech
2 business  003.txt  Yukos unit buyer faces loan claim
3 business  004.txt  High fuel prices hit BA's profits
4 business  005.txt  Pernod takeover talk lifts Domecq

                                             content
0  Quarterly profits at US media giant TimeWarne...
1  The dollar has hit its highest level against ...
2  The owners of embattled Russian oil giant Yuk...
3  British Airways has blamed high fuel prices f...
4  Shares in UK drinks and food firm Allied Dome...
```

```
[6] data.isnull().sum()
```

```
category    0
filename    0
title       0
content     0
dtype: int64
```



```
[11] df.head()
```

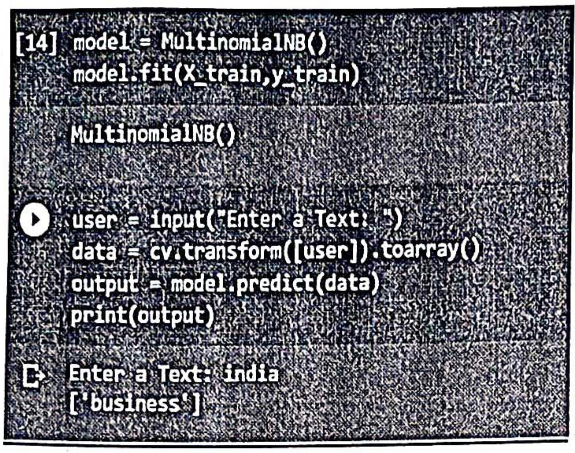
	category	filename	title	con
0	business	001.txt	Ad sales boost Time Warner profit	Quarterly profits at US media giant TimeWa
1	business	002.txt	Dollar gains on Greenspan speech	The dollar has hit its highest level agai
2	business	003.txt	Yukos unit buyer faces loan claim	The owners of embattled Russian oil giant Y
3	business	004.txt	High fuel prices hit BA's profits	British Airways has blamed high fuel pric
4	business	005.txt	Pernod takeover talk lifts Domecq	Shares in UK drinks and food firm Allied Do

```
data["category"].value_counts()
```

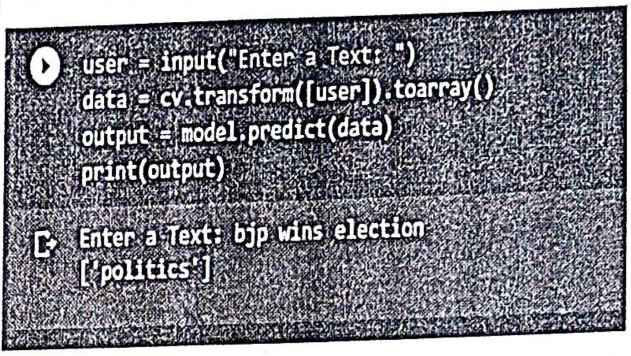
```
sport            511
business         510
politics         417
tech             401
entertainment    386
Name: category, dtype: int64
```

```
[12] data = data[["title", "category"]]
     data.head()
```

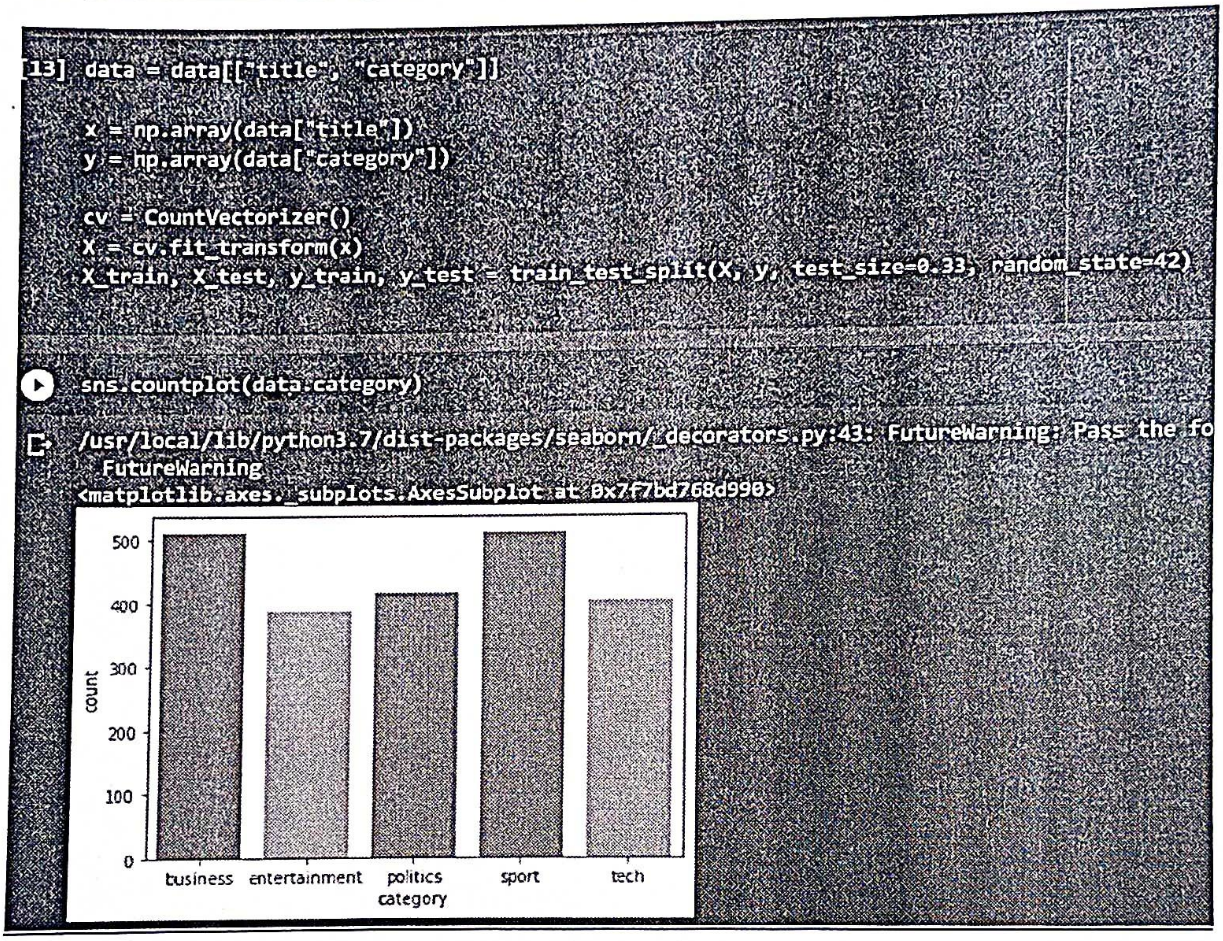
	title	category
0	Ad sales boost Time Warner profit	business
1	Dollar gains on Greenspan speech	business
2	Yukos unit buyer faces loan claim	business
3	High fuel prices hit BA's profits	business
4	Pernod takeover talk lifts Domecq	business



```
[14] model = MultinomialNB()
     model.fit(X_train,y_train)

     MultinomialNB()

     user = input("Enter a Text: ")
     data = cv.transform([user]).toarray()
     output = model.predict(data)
     print(output)

     Enter a Text: india
     ['business']
```



```
user = input("Enter a Text: ")
data = cv.transform([user]).toarray()
output = model.predict(data)
print(output)

Enter a Text: bjp wins election
['politics']
```



```
[13] data = data[["title", "category"]]

     x = np.array(data["title"])
     y = np.array(data["category"])

     cv = CountVectorizer()
     X = cv.fit_transform(x)
     X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=42)

     sns.countplot(data.category)

     /usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass the fo
       FutureWarning
     <matplotlib.axes._subplots.AxesSubplot at 0x7f7bd768d990>
```

Chapter 5 CONCLUSIONS

InthisprojectIhaveusedPythonprogramminglanguageformachinelearning.This projectcanbeusedwidelybyTVreportersandalsoforresearch purposes in the field of machine learning for segregation of data. It is under progress and is very easy to implement.

It uses multinomial naïve bayes algorithm for NLP.

SCOPE OF THE PROJECT

This machine learning model is based on the principle of supervised learning which includes classification (grouping) of data in the predefined sets. This project can be used to categories new upcoming news in the following categories hence reducing time and labor and making machine capable enough to take decisions.

It can be used widely by news channels and in research purposes.

Chapter 6 - REFERENCE AND BIBLIOGRAPHY

- https://www.javatpoint.com/machine-learning
- https://www.w3schools.com/python/python_ml_getting_started.asp
- Python.org
- https://neptune.ai/blog/vectorization-techniques-in-nlp-guide#:~:text=In%20Machine%20Learning%2C%20vectorization%20is,converting%20text%20to%20numerical%20vectors.