# Data Science

## (Project Report)

TY Computer Engineering 2021-22

Div.: 1 Batch: T4

## Group Members:

1. Nishant Badgujar (111903053)

2. Rajkumar Sawant (111903066)

3. Sanjyot Gaidhani (111903084)

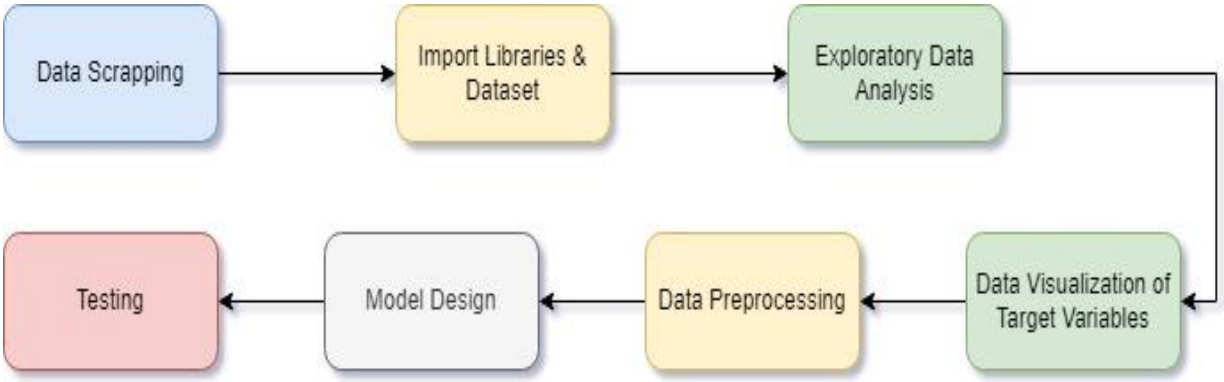
**Proposed Topic:** Twitter Sentiment Analysis for Product Review

# Introduction

Sentiment analysis refers to identifying as well as classifying the sentiments that are expressed in the text source. Tweets are often useful in generating a vast amount of sentiment data upon analysis. These data are useful in understanding the opinion of the people about a variety of Products used in day-to-day life.

# Project Pipeline

The various steps involved in the project are –




# Implementation

## Data Scraping:

- Data scraping, also known as **web scraping**, is the process of importing information from a website into a spreadsheet or local file saved on your computer.
- There are various techniques to scrap data from the web. In our project, we had scrapped tweets data using APIs provided by twitter.
- For that we need to create a Twitter Developer Account alongside the **Twint** module from Python.

## Import Necessary Dependencies Dataset:

- Various Machine Learning Libraries like **NumPy, pandas, seaborn, matplotlib** and **sklearn** should be imported there to carry out the implementation process.
- We can import the Scraped Dataset by reading and loading it from the CSV file.

## Exploratory Data Analysis:

- Exploratory Data Analysis involves exporting the five top records of data & columns/features in data.
- Also, it involves finding the length & shape of the dataset, gaining data information like Datatypes, checking for null, unique and number of target values.

## Data Visualization:

- Data visualization is the graphical representation of information and data. It is a particularly efficient way of communicating when the data is numerous as of our dataset.
- For our project we used the following visualization techniques to get information about dataset
    - Bar Plot
    - Pie Chart
    - Count Plot6

## Data Preprocessing:

- Before training the model, we must perform pre-processing steps on the dataset that deals with removing stop words and removing emojis.
- The text document is then converted into the lowercase for better generalization.
- We have also removed the repeating characters from the words along with removing the URLs as they do not have any significant importance.
    - **Stemming** (reducing the words to their derived stems)

- **Lemmatization** (reducing the derived words to their root form known as lemma) for better results.

## Model Design:

After training the model we then apply the evaluation measures to check how the model is performing. Accordingly, we use the following evaluation parameters to check the performance of the models respectively:

- Accuracy Score
- Confusion Matrix with Plot
- ROC-AUC Curve

## Model Building:

In the problem statement we have used three different models respectively:

- Bernoulli Naive Bayes
- SVM (Support Vector Machine)
- Logistic Regression

The idea behind choosing these models is that we want to try all the classifiers on the dataset ranging from simple ones to complex models and then try to find out the one which gives the best performance among them.

# Conclusion

Upon evaluating all the models, we can conclude the following details -

As far as the accuracy of the model is concerned **Support Vector Machine** performs better than **Logistic Regression** which in turn performs better than **Bernoulli Naive Bayes**.

**F1-score:** The F1 Scores for class 0 and class 1 are:
(a) For class 0: Bernoulli Naive Bayes (accuracy = 0.81) < Logistic Regression (accuracy = 0.85) < SVM (accuracy =0.89)
(b) For class 1: Bernoulli Naive Bayes (accuracy = 0.65) < Logistic Regression (accuracy = 0.70) < SVM (accuracy = 0.81)

**AUC Score:** All three models have the same **ROC-AUC** score.

We, therefore, conclude that the **SVM (Support Vector Machine)** is the best model for the above-given dataset.

# References

1] http://www.iosrjen.org/Papers/Conf.SICTIM-2019/Volume-1/5.%2022-25.pdf

2] https://www.researchgate.net/publication/343288167_Real_Time_Twitter_Sentiment_Analysis_using_Natural_Language_Processing