**PROJECT**

**Sentiment Analysis using Predictive Models**

**Problem Definition:**

We have been given restaurant details and we have to collect the data and analyse the data of that particular restaurant and work out what exactly people feel about that restaurant. Like is it a good restaurant or average restaurant or a bad restaurant using machine learning problem-solving techniques?

This analysis will help the restaurant in its future planning and growth.

**Solution Approach:**

The approach here would be like this:

* Scrapping the data
* Categorize the data based on their types like audio data, video data, and sentence data.
* Doing a deep analysis of those data.
* Modelling the data.
* Making predictions regarding the tone of data whether positive or negative or neutral.

Here is a simple flow-chart that will explain to you what my approach is towards the problem:



Let us understand this chart step by step:

* **Collecting the data**

There are various social media platforms, various websites, and various online services that act as an interface between the customer and the product (restaurant). We are going to collect data from those sources and then store that data for further analysis.

* **Categorizing data**

Generally, we have 3 types of data: Text, audio, and video.

So we are going to categorize our data separately based on the above factors and then store them further separately.

* **Cleaning of data**

There can be chances that the data that we have collected might be irrelevant or wasted, so we will clean that data 1st to get the best possible data. For cleaning, we can apply the following techniques:

* **Null value handling**- It will help us in either removing too much empty data or filling the empty data with the same relevant values.
* **Correlation techniques**- There might be chances that the data we collected might be somewhat related to each other. Here we will work on the relation of data with each other and then clean accordingly.
* **Variance methodology**: There can be chances that the data might have very low variance hence less variety, so we are going to work on those types of data and further clean them up.
* **Modelling the data**

There are some models that we can work on using our clean data.

1) Regression models

2) Classification models

Based on the datasets we are going to apply various algorithms on it for making further predictions.

* **Pre-processing the data**

Then we are going to apply NATURAL LANGUAGE PROCESSING using the NLTK method

* **Making prediction**

The final step would be to predict the data and if the data is giving us positive, negative, or neutral feedback.

**Input Data:**

We have been given the dataset, we have to just review that dataset and analyse it and then make conclusions about it.

Following are the conclusions that I made from the given dataset:

* It is of dimension 1000×2 (1000 rows and 2 columns).
* It has 2 columns Review and Liked.
* Review is of object data type.
* Liked is of integer data type.
* None of the column has null values present.

**Initial Dataset Creation:**

We have an excel file which is unclean i.e. unstructured. We need to clean the data frame and bring it to a structured form. Following are the steps that have been performed:

* Merging all the columns and forming a new column
* Removing 0 and 1 from those column and renaming the column as “New Review” column and saving the dataframe to memmory.

**Creation of our own data frame:**

I have randomly created my own 10 reviews and whether or not it is liked.

Now we will merge the above clean data frame with self-created dataset to get a whole new data frame.

**Web scrapping :**

Web scraping is an automatic method to obtain large amounts of data from websites. Most of this data is unstructured data in an HTML format which is then converted into structured data in a spread sheet or a database using some coding. This further is converted to csv or excel file.

There are many different ways to perform web scraping to obtain data from websites. These include using online services, particular API’s or even creating your code for web scraping from scratch. Many large websites, like Google, Twitter, Facebook, Stack Overflow, etc. have API’s that allow you to access their data in a structured format.

It can primarily be done using three libraries:

* Using Beautiful Soup Library
* Using Auto Scraper
* Using Selenium
* Using Parse Hub

**Extrapolatory Data Analysis(EDA):**

It is all about analysing text statistics.

They can include:

* Length of sentences per review.
* Length of words per review.

In short it is all about getting insights of the textual data. We can use various charts like bar models, histograms etc. for such type of analysis.

**Feature Engineering with Text Data:**

Feature engineering is one of the most important steps in machine learning. It is the process of using domain knowledge of the data to create features that make machine learning algorithms work.

Text themselves cannot be used by machine learning models. They expect their input to be numeric. So we need some way that can transform input text into a numeric feature in a meaningful way.

For this purpose, we perform Feature Engineering on text data. It can include:

* Counting Stop-words
* Counting Punctuations
* Upper/Lower/Tile case words or sentences.
* Number of sentences in text.
* Polarity of sentence.
* Counting parts of speech etc.

**Categorizing the Text Data:**

Now we will apply nayve bayes algo to check the categories of given data.