**“Sentiment Analysis of Amazon Musical Instruments dataset”**

Minor project report submitted in partial fulfilment of the requirement for the degree of Bachelor of Technology

in

# Computer Science and Engineering

By

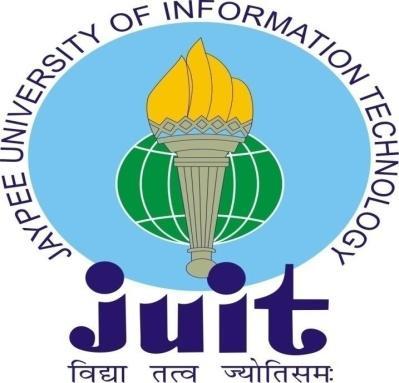
## VARDHAN CHAMBIAL (181425)

## TANVI TIWARI (181468)

**UNDER THE SUPERVISION OF**

DR. RAJNI MOHANA

(Associate Professor (Senior Grade), Department of Computer Science & Engineering), Jaypee University of Information Technology



Department of Computer Science & Engineering and Information Technology

# Jaypee University of Information Technology, Waknaghat, 173234, Himachal Pradesh, INDIA

**Candidate’s Declaration**

I hereby declare that the work presented in this project entitled **“Sentiment Analysis of Amazon Musical Instruments Reviews”**  has been done by me under the supervision of **Dr. Rajni Mohana** (Associate Professor) (Senior Grade), Department of Computer Science & Engineering), Jaypee University of Information Technology.

I also declare that the matter embodied in this project has not been submitted elsewhere for award of any degree or diploma.

**Supervised by**:

**Dr. Rajni Mohana**

Associate Professor (Senior Grade)

Department of Computer Science & Engineering and Information Technology

Jaypee University of Information Technology

**Submitted by:**

**Vardhan Chambial**, 181425  **Tanvi Tiwari**, 181468

CSE Department CSE Department

JUIT JUIT

**CERTIFICATE**

This is to certify that the work which is being presented in the project report titled **“Amazon Musical Instruments Reviews”** in partial fulfilment of the requirements for the award of the degree of B.Tech in Computer Science & Engineering and submitted to the Department of Computer Science & Engineering, Jaypee University of Information Technology, Waknaghat is an authentic record of work carried out by **Vardhan Chambial (181425)** and **Tanvi Tiwari (181468)** during the period from January 2021 to May 2021 under the supervision of **Dr. Rajni Mohana**, Department of Computer Science and Engineering, Jaypee University of Information Technology, Waknaghat.

Vardhan Chambial Tanvi Tiwari

(181425) (181468)

The above statement made is correct to the best of my knowledge.

Dr. Rajni Mohana

Associate Professor (Senior Grade)

Department of Computer Science & Engineering and Information Technology

Jaypee University of Information Technology

**ACKNOWLEDGEMENT**

Firstly, I express my heartiest thanks and gratefulness to almighty God for His divine blessing makes it possible to complete the project work successfully.

I am really grateful and wish my profound indebtedness to my Supervisor **Dr. Rajni Mohana**, Associate Professor (Senior Grade), Department of CSE Jaypee University of Information Technology, Wakhnaghat. Deep Knowledge & keen interest of my supervisor in the field of “**Sentiment Analysis**” to carry out this project. Her endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stages have made it possible to complete this project.

I would like to express my heartiest gratitude to **Dr. Rajni Mohana,** Department of CSE, for his kind help to finish my project.

I would also generously welcome each one of those individuals who have helped me straightforwardly or in a roundabout way in making this project a win. In this unique situation, I might want to thank the various staff individuals, both educating and non-instructing, which have developed their convenient help and facilitated my undertaking.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

Vardhan Chambial Tanvi Tiwari

(181425) (181468)

**TABLE OF CONTENTS**

|  |  |
| --- | --- |
| *Content* | *Page No.* |
| *Declaration by Candidate* | *II* |
| *Certificate by Supervisor* | *III* |
| *Acknowledgement* | *IV* |
| *Abstract* | *VI* |

**INDEX**

|  |  |
| --- | --- |
| CHAPTER 1 | INTRODUCTION |
| * 1. Introduction   2. Objective   3. Motivation   4. Language Used   1.5 Project Deliverables | |  |
| CHAPTER 2 | | **MINOR PROJECT SLDC** |
| 2.1 Feasibility Study on Minor Project   * 1. Requirements on Minor Project   2.2.1 Functional Requirements  2.2.2 Non-Functional Requirements  2.3 Use Case Diagram of the Project  2.4 DFD Diagram of the Project  2.5 State Transition Diagram of Project | |  |
| CHAPTER 3 | | **PROJECT IMPLEMENTATION** |
| 3.1 Date Set Used in the Minor Project   * 1. Date Set Features      1. Types of Data Set      2. Features of the data set   3.3 Design of Problem Statement  3.4 Algorithm / Pseudo code of Project Problem  3.5 Flow Graph of the Project  3.6 Screenshots | |  |
| CHAPTER 4 | | **RESULTS** |
| 4.1 Discussion on the Results Achieved  4.2 Application of the Project  4.3 Limitation of the Project  4.4 Future Work | |  |
| REFERENCES | | |

**ABSTRACT**

The web could be a giant virtual house whereby to precise and to share individual

evaluations, influencing something of life, with implications for advertising and voice communication alike. Social media area unit influencing clients’ potentialities by the approach of shaping their attitudes and behaviors. Chase the social media activities could be a great way to live customers’ loyalty, retentive a tune on their sentiment towards brands or merchandise. Social media area unit the subsequent logical advertising space.

Web Portals get a vast amount of feedback from the users. To go through all the feedback can be a tedious job. You have to categorize opinions expressed in feedback forums. This can be utilized for a feedback management system. We classify individual comments/reviews.and we also determine overall rating based on individual comments/reviews. So that company can get a complete idea on feedback provided by the customers and can take care of those particular fields. This makes more loyal Customers to the company, increase in business , fame, brand value, profits.

**Sentiment Analysis** is a Language-independent technology that understands the meaning of the text. It identifies the opinion or attitude that a person has towards a topic or an object. Sentiment analysis is widely applied to reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine.

With the help of Sentimental Analysis, we tend to help the organization to understand better about their customer feedback so that they can concentrate on the issues faced by the customers.

**Chapter 01: INTRODUCTION**

**1.1 Introduction**

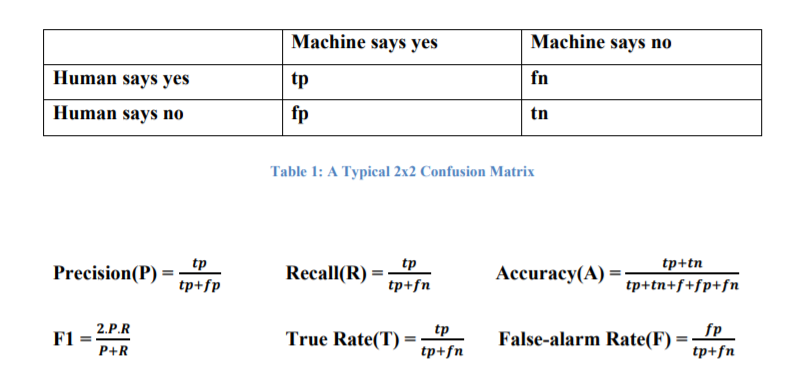
Sentiment is an attitude, thought, or judgment prompted by feeling. Sentiment analysis, which is also known as opinion mining, studies people’s sentiments towards certain entities. From a user’s perspective, people are able to post their own content through various social media, such as forums, micro-blogs, or online social networking sites. From a researcher’s perspective, many social media sites release their application programming interfaces (APIs), prompting data collection and analysis by researchers and developers. However, those types of online data have several flaws that potentially hinder the process of sentiment analysis. The first flaw is that since people can freely post their own content, the quality of their opinions cannot be guaranteed. The second flaw is that the ground truth of such online data is not always available. A ground truth is more like a tag of a certain opinion, indicating whether the opinion is positive, negative, or neutral.

As online marketplaces have been popular during the past decades, the online sellers and merchants ask their purchasers to share their opinions about the products they have bought. Everyday millions of reviews are generated all over the Internet about different products, services and places. This has made the Internet the most important source of getting ideas and opinions about a product or a service. However, as the number of reviews available for a product grows, it is becoming more difficult for a potential consumer to make a good decision on whether to buy the product. Different opinions about the same product on one hand and ambiguous reviews on the other hand makes customers more confused to get the right decision. Here the need for analyzing this content seems crucial for all e-commerce businesses.

Sentiment analysis and classification is a computational study which attempts to address this problem by extracting subjective information from the given texts in natural language, such as opinions and sentiments. Different approaches have been used to tackle this problem from natural language processing, text analysis, computational linguistics, and biometrics. In recent years, Machine learning methods have become popular in the semantic and review analysis for their simplicity and accuracy. Amazon is one of the e-commerce giants that people are using every day for online purchases where they can read thousands of reviews dropped by other customers about their desired products. These reviews provide valuable opinions about a product such as its property, quality and recommendations which helps the purchasers to understand almost every detail of a product. This is not only beneficial for consumers but also helps sellers who are manufacturing their own products to understand the consumers and their needs better.

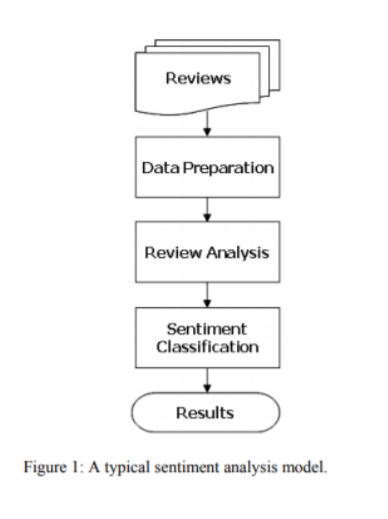
This project is considering the sentiment classification problem for online Music Instrument reviews using supervised approaches to determine the overall semantic of customer reviews by classifying them into positive and negative sentiment. Classification techniques can also be divided into two categories: Supervised vs. unsupervised and non-adaptive vs. adaptive/reinforcement techniques. Supervised approach is when we have pre-labeled data samples available and we use them to train our classifier. Training the classifier means to use the pre-labeled to extract features that best model the patterns and differences between each of the individual classes, and then classifying an unlabeled data sample according to whichever pattern best describes it. Unsupervised classification is when we do not have any labeled data for training. In addition to this adaptive classification techniques deal with feedback from the environment.

There are several metrics proposed for computing and comparing the results of our experiments. Some of the most popular metrics include: Precision, Recall, Accuracy, F1-measure, True rate and False alarm rate (each of these metrics is calculated individually for each class and then averaged for the overall classifier performance.) A typical confusion table for our problem is given below along with an illustration of how to compute our required metric.



**1.2 Project Objective**

The objective of this project is to analyze the ***Musical Instrument*** reviews dataset which we have taken from Kaggle and perform Sentiment Analysis on it in order to predict whether the people like the product or they don’t.



* Reviews Preprocessing and Cleaning
* Story Generation and Visualization from reviews
* Extracting Features from Cleaned reviews
* Model Building: Sentiment Analysis

The algorithm and method used in the development of the project should be efficient and at the same time flexible so that one can execute it. The system should not be rigid and platform dependent. The system should be user-friendly and accurate.

**1.3 Project Motivation**

Every day we come across various products in our lives, on the digital medium we swipe across hundreds of product choices under one category. It will be tedious for the customer to make a selection. Here comes 'reviews' where customers who have already got that product leave a rating after using them and brief their experience by giving reviews. As we know ratings can be easily sorted and judged whether a product is good or bad. But when it comes to sentence reviews we need to read through every line to make sure the review conveys a positive or negative sense. In the era of artificial intelligence, things like that have got easier with the Natural Language Processing (NLP) technology. Also the fact that we can make estimations, predictions and give the ability for machines to learn by themselves is both powerful and limitless in terms of application possibilities.

Motivation for Sentiment Analysis is two-fold. Both consumers and producers highly value “customer’s opinion” about products and services.

**The Consumer’s Perspective :**

While taking a decision it is very important for us to know the opinion of the people around us. Earlier this group used to be small, with a few trusted friends and family members. But, now with the advent of the Internet we see people expressing their opinions in blogs and forums. These are now actively read by people who seek an opinion about a particular entity (product, movie etc.). Trying to go through such a vast amount of information to understand the general opinion is impossible for users just by the sheer volume of this data. Hence, the need for a system that differentiates between good reviews and bad reviews. Further, labeling these documents with their sentiment would provide a succinct summary to the readers about the general opinion regarding an entity.

**The Societies’ Perspective :**

Recently, certain events, which affected the Government, have been triggered using the Internet. The social networks are being used to bring together people so as to organize mass gatherings and oppose oppression. On the darker side, the social networks are being used to insinuate people against an ethnic group or class of people, which has resulted in a serious loss of life. Thus, there is a need for Sentiment Analysis systems that can identify such phenomena and curtail them if needed.

**1.4 Language Used**

**Python:**

Python is a deciphered, object-arranged, raised level programming with enthusiastic

semantics. Its raised level certain data structures, got together with exuberant creating

and sure, make it engaging for Rapid Application Development, similarly concerning

use as scripting or glue language to relate existing parts together. Python's clear, easy

to learn language structure underlines unequivocal quality and appropriately

diminishes the cost of the program support. It supports parts and packages, which

invigorates program protection and code reuse.

**Machine Learning :**

Machine learning empowers PCs to learn without being explicitly tweaked (Arthur Samuel, 1959).This is a subfield of programming building. Machine learning ventures to every part of the advancement of estimations which can learn and make possibilities on data. Such controls stick to altered rules, yet can in like manner choose forecasts or decisions subject to data. They build models from test inputs. Machine learning is done where masterminding and programming express computations is unthinkable. Models incorporate spam filtering.

**Natural Language Processing :**

Sentiment analysis is a natural language processing (NLP) technique that’s used to classify subjective information in text or spoken human language. The objective of sentiment analysis is to categorize the sentiment of public opinions by sorting them into positive, neutral, and negative. And ***Python*** is often used in NLP tasks like sentiment analysis because there are a large collection of NLP tools and libraries to choose from. There are many things about Python that make it a really good programming language choice for an NLP project. The simple syntax and transparent semantics make it an excellent choice for projects that include Natural Language Processing tasks.

There’s something else about this versatile language Python that makes it such a great technology for helping machines process natural languages. It provides collection of NLP tools and libraries that enable developers to handle a great number of NLP-related tasks such as document classification, topic modeling, part-of-speech (POS) tagging, word vectors, and sentiment analysis.

**Natural Language Toolkit (NLTK) :**

NLTK is an essential library that supports tasks such as classification, stemming, tagging, parsing, semantic reasoning, and tokenization in Python. It’s basically your main tool for natural language processing and machine learning. Today it serves as an educational foundation for Python developers who are dipping their toes in this field. This library is pretty versatile, but we must admit that it’s also quite difficult to use for Natural Language Processing with Python..

Today, Python is one of the most popular programming languages for this task and it has replaced many languages in the industry, one of the reasons is its vast collection of libraries. Python libraries that used in Machine Learning are:

* Numpy
* Scipy
* Pandas
* Matplotlib
* Seaborn

**1.5 Technical Requirements (Hardware )**

Ram : 1 GB Ram and more

Processor : Any Intel Processor

HardDisk : 6 GB and more

Speed : 1 GHZ and more

* Linux Operating System/Windows
* Modern Web Browser
* Python Platform (Anaconda2, Spyder, Jupyter)
* NLTK Toolkit

**1.6 Deliverables of the Project**

* Web Portals get a vast amount of feedback from the users. To go through all the feedback can be a tedious job. We have to categorize opinions expressed in feedback forums.
* This can be utilized for a feedback management system. We classify individual comments/reviews and we also determine overall rating based on individual comments/reviews.
* So the company can get a complete idea on feedback provided by the customers and can take care of those particular fields. This makes more loyal Customers to the company, increase in business , fame, brand value, profits.
* Our project emphasizes more on buyers’ and helps them to have a better idea about any Musical Instrument that they are interested in just by having a wholesome look on the reviews given by people.

**Chapter 02:PROJECT SOFTWARE DEVELOPMENT LIFE CYCLE**

**2.1 Feasibility Study on Minor Project**

Feasibility Study in Software Engineering is a study to evaluate feasibility of a proposed project. As the name suggests, a feasibility study is the feasibility analysis or it is a measure of the software product in terms of how beneficial product development will be for the organization in a practical point of view. Feasibility study is carried out based on many purposes to analyze whether a software product will be right in terms of development, implantation, contribution of project to the organization etc. Feasibility study is an important stage of Software Project Management Process as after completion of feasibility study it gives a conclusion of whether to go ahead with proposed project as it is practically feasible or to stop proposed project here as it is not right/feasible to develop or to think/analyze about proposed project again. Along with this Feasibility study helps in identifying risk factors involved in developing and deploying systems and planning for risk analysis also narrows the business alternatives and enhances success rate analyzing different parameters associated with proposed project development.

The main objectives of the feasibility study are to determine whether the project would be feasible in terms of economic feasibility, technical feasibility and operational feasibility and schedule feasibility or not. It is to make sure that the input data which are required for the project are available. Thus we evaluated the feasibility of the system in terms of the following categories: Technical feasibility Operational feasibility Economic feasibility Schedule feasibility

**Types of Feasibility Study :**

* **Technical Feasibility :**

Evaluating the technical feasibility is the trickiest part of a feasibility study. This is because, at the point in time there is no detailed design of the system, making it difficult to access issues like performance, costs (on account of the kind of technology to be deployed) etc. A number of issues have to be considered while doing a technical analysis; understand the different technologies involved in the proposed system. Before commencing the project, we have to be very clear about what are the technologies that are to be required for the development of the new system. Is the required technology available? Our system is technically feasible since all the required tools are easily available. Google Colab and Python makes the system more user and developer friendly and although all tools seem to be easily available.

* **Operational Feasibility :**

In Operational Feasibility the degree of providing service to requirements is analyzed along with how easy the product will be to operate and maintain after deployment. Along with this other operational scopes are determining usability of the product, Determining suggested solutions by software development team is acceptable or not etc. Proposed project is beneficial only if it can be turned into information systems that will meet the operating requirements. Simply stated, this test of feasibility asks if the system will work when it is developed and installed. Are there major barriers to Implementation? The proposal was based on the Feedback Management System in which we have to do Sentiment Analysis on Amazon musical instrument reviews. It is simple though there are challenges too. It is free and not costly to operate.

* **Economic Feasibility :**

In the Economic Feasibility study cost and benefit of the project is analyzed. Means under this feasibility study a detailed analysis is carried out on the cost of the project for development which includes all required costs for final development like hardware and software resource required, design and development cost and operational cost and so on. After that it is analyzed whether the project will be beneficial in terms of finance for organization or not.

A simple economic analysis which gives the actual comparison of costs and benefits is much more meaningful in this case. In addition, this proves to be a useful point of reference to compare actual costs as the project progresses. There could be various types of intangible benefits on account of automation. These could increase improvement in product quality, better decision making, and timeliness of information, expediting activities, improved accuracy of operations, better documentation and record keeping, faster retrieval of information. Here, sentiment analysis of amazon musical instruments is to be done using Machine Learning techniques and is not costly.

* **Legal Feasibility :**

The Legal Feasibility study project is analyzed from a legality point of view. This includes analyzing barriers of legal implementation of project, data protection acts or social media laws, project certificate, license, copyright etc. Overall it can be said that the Legal Feasibility Study is a study to know if proposed projects conform to legal and ethical requirements. Here , our project is completely legal following all the required norms.

* **Schedule Feasibility :**

In Schedule Feasibility Study mainly timelines/deadlines are analyzed for proposed projects which includes how many times teams will take to complete the final project which has a great impact on the organization as the purpose of the project may fail if it can’t be completed on time. Here, our project has been done within the time frame and ready to be in use.

**2.2 Requirements on Minor Project**

After the extensive analysis of the problems in the system, we are familiarized with the requirements that the current system needs. The requirement that the system needs is categorized into the functional and non-functional requirements. These requirements are listed below:

**2.2.1 Functional Requirements :**

Functional Requirements are the requirements that the end user specifically demands as basic facilities that the system should offer. These are represented in the form of input to be given to the system, the operation performed and the output expected. They are basically the requirements stated by the user which one can see directly in the final product, unlike the non-functional requirements. The features that must be included in any system to satisfy the business needs and be acceptable to the users.

Based on this, the functional requirements that the system must require are as follows:

* System should be able to process new reviews stored in the database after retrieval.
* System should be able to analyze data and classify each review’s sentiment.

**2.2.2 Non-Functional Requirements :**

Non-functional requirements is a description of features, characteristics and attributes of the system as well as any constraints that may limit the boundaries of the proposed system. The non-functional requirements are essentially based on the performance, information, economy, control and security efficiency and services. They basically deal with issues like Portability, Security, Maintainability, Reliability, Scalability, Performance, Reusability and Flexibility.

Based on these the non-functional requirements of our project are as follows:

* User friendly
* System should provide better accuracy
* To perform with efficient throughput and response time.

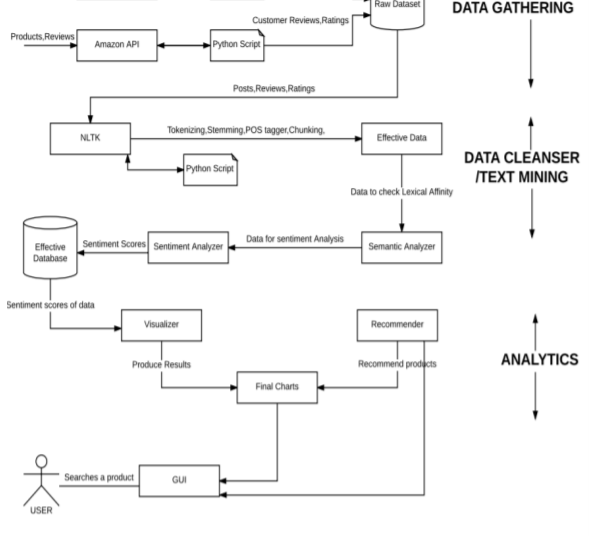
**2.3 Use Case Diagram of the Project**

A use case diagram is used to represent the dynamic behavior of a system. It encapsulates the system's functionality by incorporating use cases, actors, and their relationships. It models the tasks, services, and functions required by a system/subsystem of an application. It depicts the high-level functionality of a system and also tells how the user handles a system.

A use case is a methodology used in system analysis to identify, clarify, and organize system requirements. In this context, the term "system" refers to something being developed or operated, such as a mail order product sales and service website. Use case diagrams are employed in UML (Unified Modeling Language), a standard notation for the modeling of real-world objects and systems.

*A use case diagram contains four components.*

* The boundary, which defines the system of interest in relation to the world around it.
* The actors, usually individuals involved with the system, are defined according to their roles.
* The use cases, which the specific roles are played by the actors within and around the system.
* The relationships between and among the actors and the use cases diagram.



***: Use Case Diagram :***

**2.4 DFD Diagram of the Project**

A Data Flow Diagram (DFD) is a traditional visual representation of the information flows within a system. A neat and clear DFD can depict the right amount of the system requirement graphically. It can be manual, automated, or a combination of both.

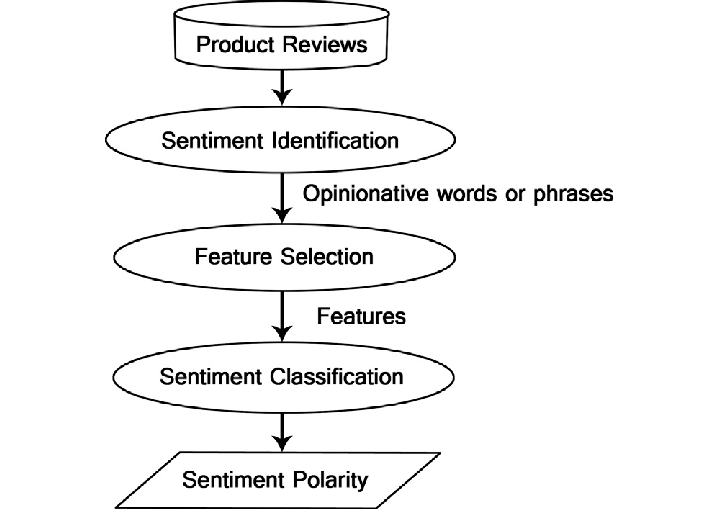
It shows how data enters and leaves the system, what changes the information, and where data is stored.

The objective of a DFD is to show the scope and boundaries of a system as a whole.

The DFD is also called a data flow graph or bubble chart.

DFD graphically represents the functions, or processes, which capture, manipulate, store, and distribute data between a system and its environment and between components of a system. The visual representation makes it a good communication tool between User and System designer. DFD has often been used due to the following reasons:

* Logical information flow of the system
* Determination of physical system construction requirements
* Simplicity of notation
* Establishment of manual and automated systems requirements

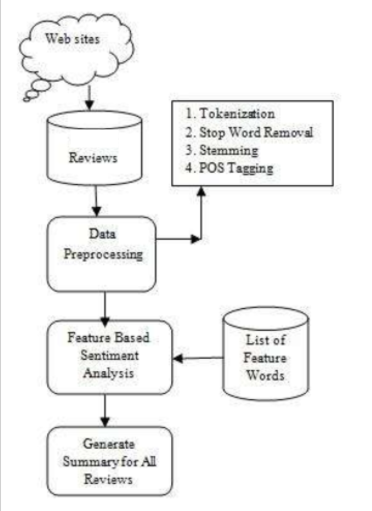


***: Data Flow Diagram of Project:***

**2.5 State Transition Diagram of the Minor Project**

State-transition diagrams describe all of the states that an object can have, the events under which an object changes state (transitions), the conditions that must be fulfilled before the transition will occur (guards), and the activities undertaken during the life of an object (actions). State-transition diagrams are very useful for describing the behavior of individual objects over the full set of use cases that affect those objects. State-transition diagrams are not useful for describing the collaboration between objects that cause the transitions. Uses of statechart diagram –

* We use it to state the events responsible for change in state (we do not show what processes cause those events).
* We use it to model the dynamic behavior of the system .
* To understand the reaction of objects/classes to internal or external stimuli.

******

***: Block Diagram of Project :***

**Chapter 03: PROJECT IMPLEMENTATION**

**3.1 Date Set Used in the Minor Project**

The Amazon reviews dataset consists of Musical Instrument reviews from Amazon.

The dataset is available on Kaggle, we have processed it from Kaggle itself. This dataset was created by Chetan Gadge. It contains the following files: Reviews include reviewer ID , User ID, Reviewer Name, Reviewer text, helpful, Summary(obtained from Reviewer text),Overall Rating on a scale 5, Review time. For more information, please refer to the Problem Statement given by ISRO to classify the customer comments. This would be helpful for the organization to understand Customer feedbacks.

"Small" subsets for experimentation are considered before requesting the larger files.

**K-cores** (i.e., dense subsets): These data have been reduced to extract the k-core, such that each of the remaining users and items have k reviews each.

**Ratings only:** These datasets include no metadata or reviews, but only (user, item, rating, timestamp) tuples. Thus they are suitable for use with mymedialite (or similar) packages.

The Amazon reviews full score dataset is constructed by randomly taking 10,000 samples for each review score from 1 to 5. In total there are 1,00,000 samples.

The format of the downloaded file was one-review-per-line in JSON. The file was converted to the Comma Separated Values (CSV) format, as it is more convenient for python to handle this type of file.

**DATA FORMAT:**

The following is an example review in Json file:

**"reviewText":** string "Not much to write about here, but it does exactly what it's supposed to. filters out the pop sounds. now my recordings are much more crisp. it is one of the lowest prices pop filters on amazon so might as well buy it, they honestly work the same despite their pricing,"

**"overall":** int 5

**"summary":** string "good"

**"unixReviewTime":** int 1393545600

**"reviewTime":** string "02 28, 2014"

**3.2 Date Set Features**

**3.2.1 Types of Data Set**

A Data set is a set or collection of data. This set is normally presented in a tabular pattern. Every column describes a particular variable. And each row corresponds to a given member of the data set. Data sets describe values for each variable for unknown quantities of an object or values of random numbers. The values in this set are known as a datum. The data set consists of data of one or more members corresponding to each row. A data set is an ordered collection of data. While handling the data, the data set can be a bunch of tables, schema and other objects. The data are essentially organized to a certain model that helps to process the needed information. The set of data is any permanently saved collection of information that usually contains either case-level, gathered data, or statistical guidance level data.

*Types of Data Sets :*

In Statistics, we have different types of data sets available for different types of information. They are:

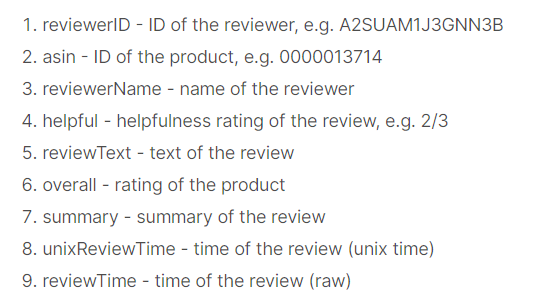
* **Numerical Data Sets :** A set of all numerical data. It deals only with numbers.
* **Bivariate Data Sets :** A data set that has two variables is called a Bivariate data set. It deals with the relationship between the two variables.
* **Multivariate Data Sets :** A data set with multiple variables.
* **Categorical Data Sets :** Categorical data sets represent features or characteristics of a person or an object.
* **Correlation Data Sets :** The set of values that demonstrate some relationship with each other indicates correlation data sets. Here the values are found to be dependent on each other.

Here, our dataset **Amazon Musical Instruments Reviews** from Kaggle is Multivariate dataset.

**3.2.2 Features of the data set**

The Amazon reviews full score dataset is constructed by randomly taking 10,000 samples for each review score from 1 to 5. In total there are 1,00,000 samples.

The format of the downloaded file was one-review-per-line in JSON. The file was converted to the Comma Separated Values (CSV) format, as it is more convenient for python to handle this type of file.



***: Describes features of each Instrument :***

**3.3 Design of Problem Statement**

We have to perform Sentiment Analysis for this data. In order to understand better about the customer feedback. So that they can concentrate on those issues customers are facing. Build the model which has highest accuracy in classifying the feedback as positive,Negative and neutral.

**Task :1**

We have to categorize opinions expressed in feedback forums

**Task :2**

We have to classify individual comments/reviews and you have to determine overall rating based on individual comments/reviews.

**3.4 Algorithm / Pseudo code of the Project Problem**

**Data preparation**

For preparing the desired data, a simple code was written in python to remove the useless features. Many features were removed except overall, rtexts, senti\_score, sentiment. The score that is generated by the reviewer includes a number of ratings on scales of 1 to 5. Reviews that were rated with one, two or three were considered as negative and those with four or five numbers were considered as positive. Reviews with three stars usually contain many mixed reviews and are difficult to be labeled into a positive or negative category.

In this study, two tasks have been performed.

In the first task the whole data set was used. Since the number of reviews were quite enough to get a reasonable result from the classifiers, the reviews with three stars were omitted to avoid any complication while training the algorithms.

However, in the second task due to the small number of data the reviews with three stars were also considered as negative. The same code was then used to label the data. The reviews that were considered as positive got a score of ”1” and the remaining ones got a ”0” score.

**SENTIMENT SENTENCE EXTRACTION & POS TAGGING:**

Tokenization of reviews after removal of STOP words which mean nothing related to sentiment is the basic requirement for POS tagging. After proper removal of STOP words like “am, is, are, the, but” and so on the remaining sentences are converted into tokens. These tokens take part in POS tagging.

In natural language processing, part-of-speech (POS) taggers have been developed to classify words based on their parts of speech. For sentiment analysis, a POS tagger is very useful because of the following two reasons:

* Words like nouns and pronouns usually do not contain any sentiment. It is able to filter out such words with the help of a POS tagger.
* A POS tagger can also be used to distinguish words that can be used in different parts of speech.

**LEMMATIZATION**

Has the goal of lessening a word to its base structure and gathering various types of a similar word. For instance, action words in past tense are changed into present (for example "went" is changed to "go") and equivalents are bound together (for example "best" is changed to "acceptable"), subsequently normalizing words with comparable importance to their root. Despite the fact that it appears to be firmly identified with the stemming procedure, lemmatization utilizes an alternate way to deal with arriving at the root types of words. Lemmatization settles words to their word reference structure (known as lemma) for which it requires itemized word references in which the calculation can investigate and interface words to their comparing lemmas.

**SENTIMENT CLASSIFICATION ALGORITHMS**

**Logistic Regression**

Logistic regression predicts the probability of an outcome that can only have two values (i.e. a dichotomy). The prediction is based on the use of one or several predictors (numerical and categorical). A linear regression is not appropriate for predicting the value of a binary variable for two reasons :

* A linear regression will predict values outside the acceptable range (e.g. predicting probabilities outside the range 0 to 1)
* Since the dichotomous experiments can only have one of two possible values for each experiment, the residuals will not be normally distributed about the predicted line.

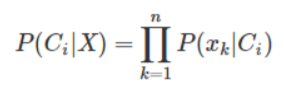
On the other hand, a logistic regression produces a logistic curve, which is limited to values between 0 and 1. Logistic regression is similar to a linear regression, but the curve is constructed using the natural logarithm of the “odds” of the target variable, rather than the probability. Moreover, the predictors do not have to be normally distributed or have equal variance in each group.

Logistic regression uses maximum likelihood estimation (MLE) to obtain the model coefficients that relate predictors to the target. After this initial function is estimated, the process is repeated until LL (Log Likelihood) does not change significantly.

**Naïve Bayesian Classifier:**

The Naïve Bayesian classifier works as follows:

Suppose that there exists a set of training data, D, in which each tuple is represented by an n-dimensional feature vector, X=x 1,x 2,..,x n , indicating n measurements made on the tuple from n attributes or features. Assume that there are m classes, C 1,C 2,...,C m . Given a tuple X, the classifier will predict that X belongs to C i if and only if: P(C i |X)>P(C j |X), where i,j∈[1,m]a n d i≠j. P(C i |X) is computed as:

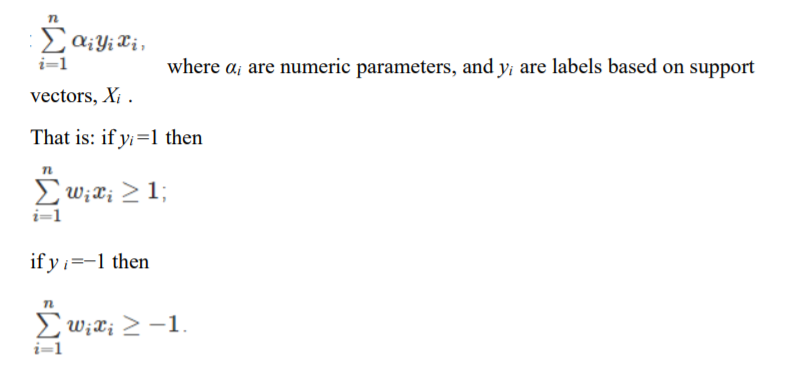
****

**Support Vector Machine (SVM)**

It is a method for the classification of both linear and nonlinear data. If the data is linearly separable, the SVM searches for the linear optimal separating hyperplane (the linear kernel),

which is a decision boundary that separates data of one class from another. Mathematically, a separating hyperplane can be written as: W·X+b=0, where W is a weight vector and W=w1,w2,...,w n. X is a training tuple. b is a scalar. In order to optimize the hyperplane,

, the problem essentially transforms to the minimization of ∥W∥, which is eventually computed as:



**Decision Tree**

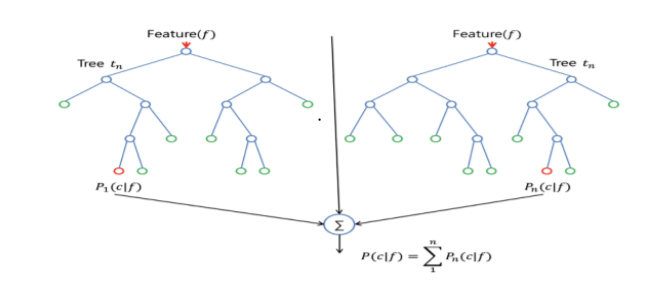
It is a classifier model in which every hub/node of the tree speaks to a test on the attribute of a data set, and its children speak to the results/outcomes. The leaf node speaks to the final classes of the data set. It is a supervised classifier (model) that utilizes data with realized names (labels) to prepare the decision tree and afterward the model is applied to the test information (test data set). For every node in the tree, the best test condition or decision has P to be taken. Utilize the GINI factor to choose the best split. For a given hub (node) t, where ( | ) is the general recurrence (Relative Frequency) of class j at node t.

GINI (t) = 1 -∑j [P (j | t)]

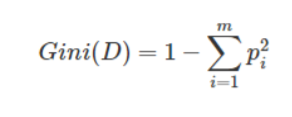
The main use of the DecisionTree Classifier from sklearn.tree package provided by scikit-learn is to build a model. GINI is used to evaluate the split at every node and the best split is always chosen. The model performed slightly better using the presence feature compared to frequency. Also using unigrams with or without bigrams didn’t make any significant improvements.

**Random forest**

The random forest classifier was chosen due to its superior performance over a single decision tree with respect to accuracy. It is essentially an ensemble method based on bagging. The classifier works as follows: Given D, the classifier firstly creates k bootstrap samples of D, with each of the samples denoted as Di . A Di has the same number of tuples as D that are sampled with replacement from D. By sampling with replacement, it means that some of the original tuples of D may not be included in Di , whereas others may occur more than once. The classifier then constructs a decision tree based on each Di.

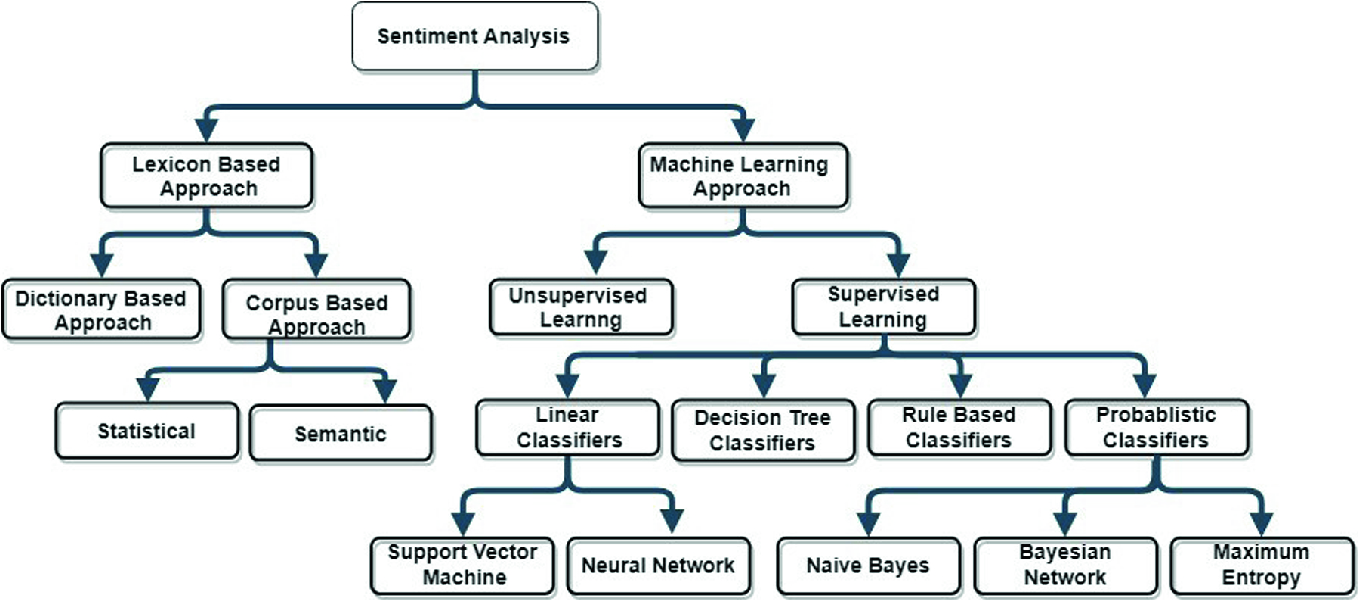
****

As a result, a “forest" that consists of k decision trees is formed. To classify an unknown tuple, X, each tree returns its class prediction counting as one vote. The final decision of X’s class is assigned to the one that has the most votes. The decision tree algorithm implemented in scikit-learn is CART (Classification and Regression Trees). CART uses Gini index for its tree induction. For D, the Gini index is computed as:



Where pi is the probability that a tuple in D belongs to class Ci . The Gini index measures the impurity of D. The lower the index value is, the better D was partitioned.

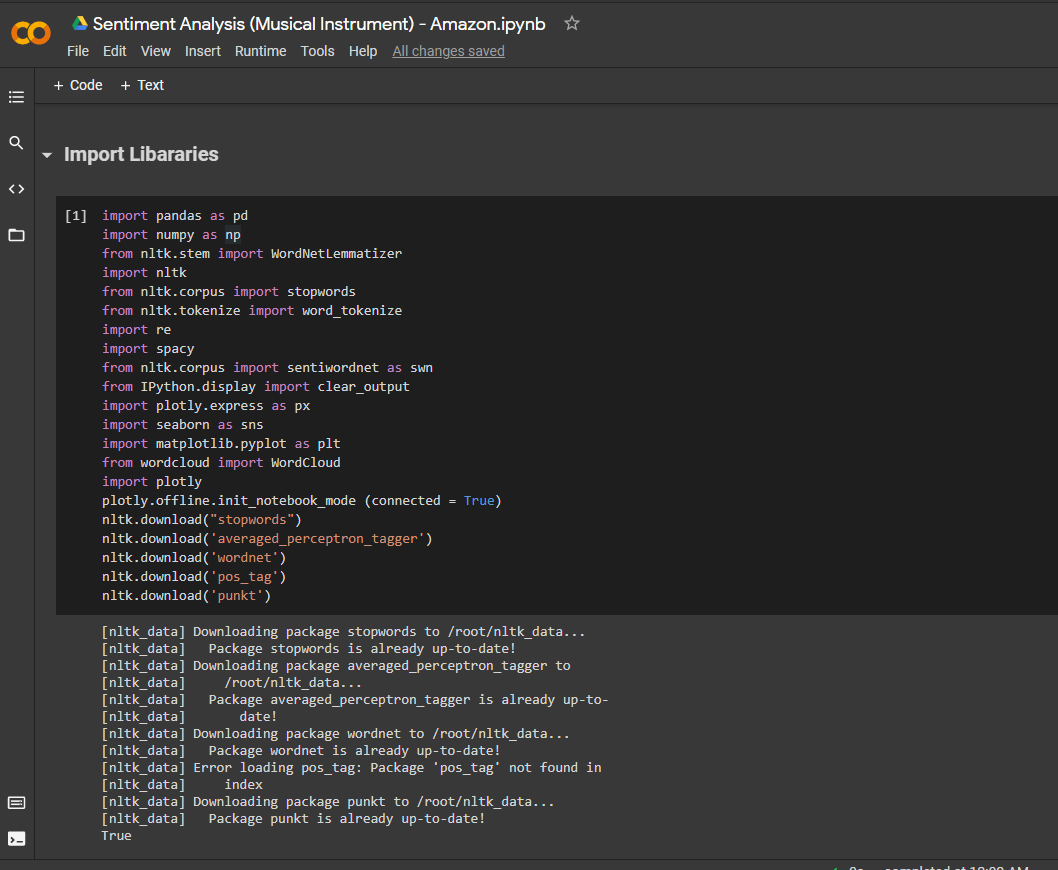
**3.5 Flow Graph of the Project**

****

***: Flow Chart Semantics :***

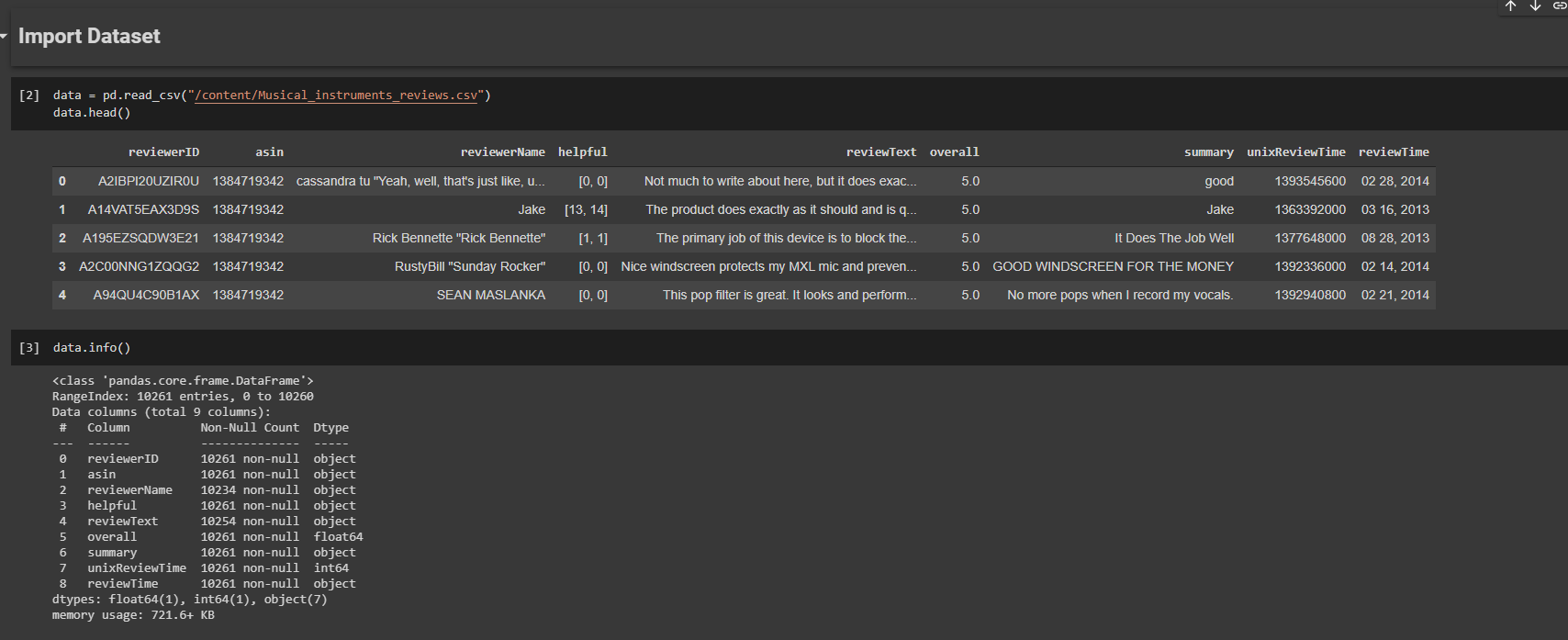
**3.6 Screenshots of the various stages of the Project**

***Analyze the Basic Structure :***

****

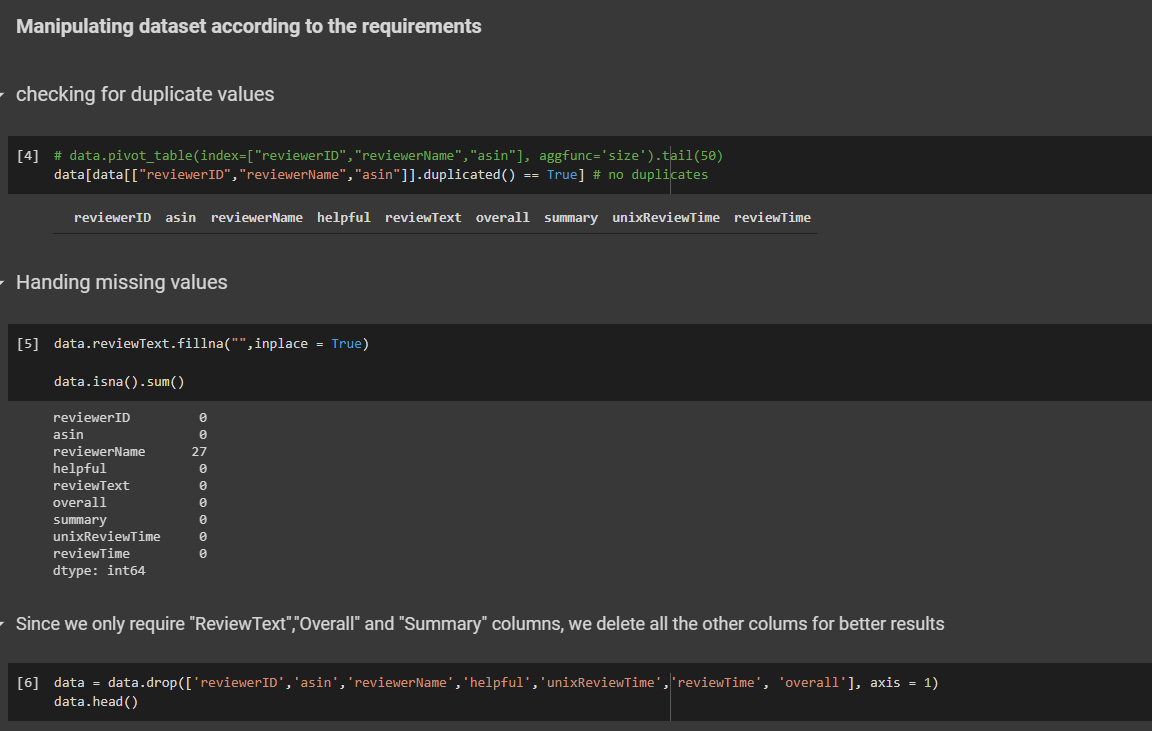
***: Basic Structure :***

***Dataset :***

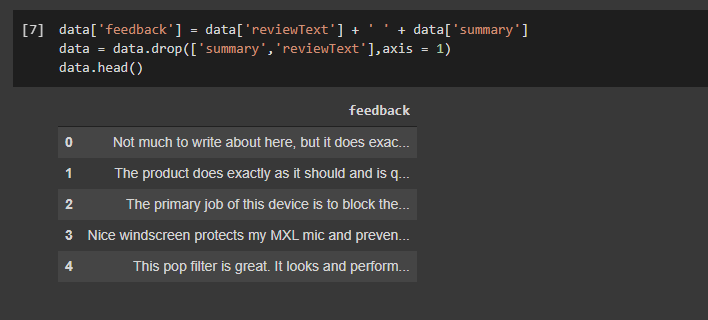
****

***: Importing Dataset :***

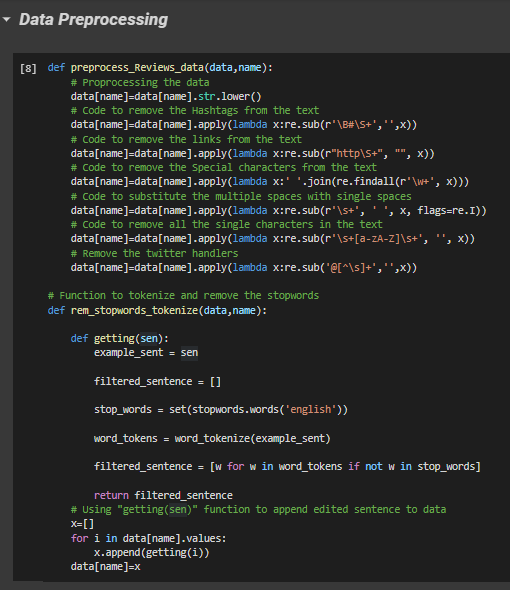
***Here we check for duplicate values as well as handling of missing values :***

****

***: Data Manipulation :***

****

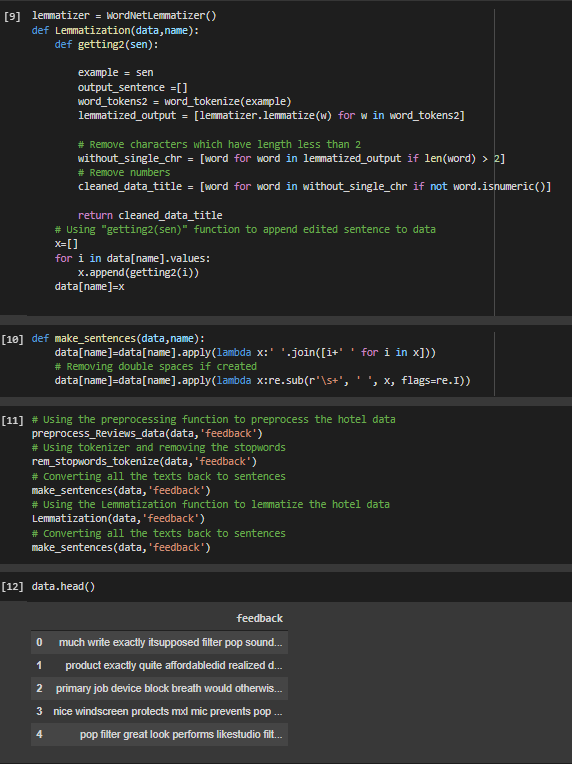
***Data Preprocessing :***

****

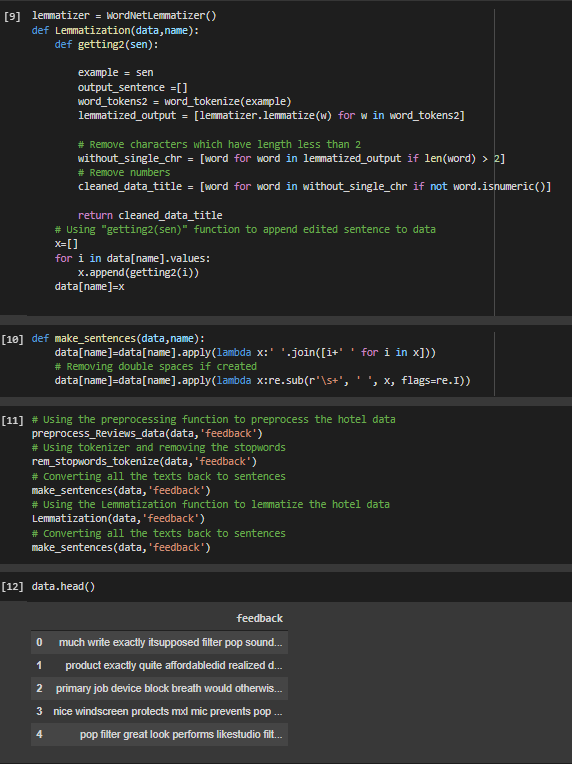
This is a vital part of training the dataset. Here Words present in the file are accessed both as a solo word and also as pair of words. Because, for example the word “bad” means negative but when someone writes “not bad” it refers to as positive. In such cases considering single word for training data will work otherwise. So words in pairs are checked to find the occurrence to modifiers before 19 any adjective which if present which might provide a different meaning to the outlook.

* The Accuracy, Precision, Recall, and Evaluation time is calculated and displayed.
* Navie Bayes, Logistic Regression, Multi SVM and Random forest, Decision Tree classifiers are applied on the dataset for evaluation of sentiments.
* Prediction of test data is done and Confusion Matrix of prediction is displayed.
* Total positive and negative reviews are counted.
* A review like sentence is taken as input on the console and if positive the console gives 1 as output and 0 for negative input.

***Lemmatization :***

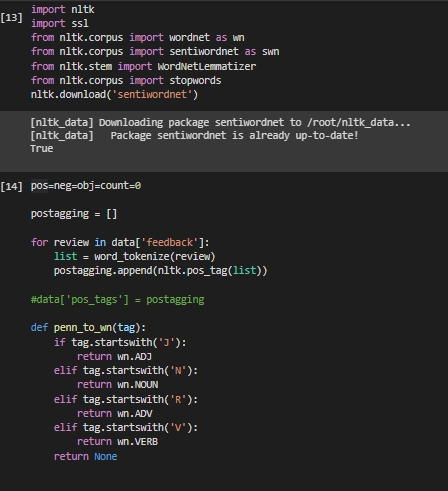
****

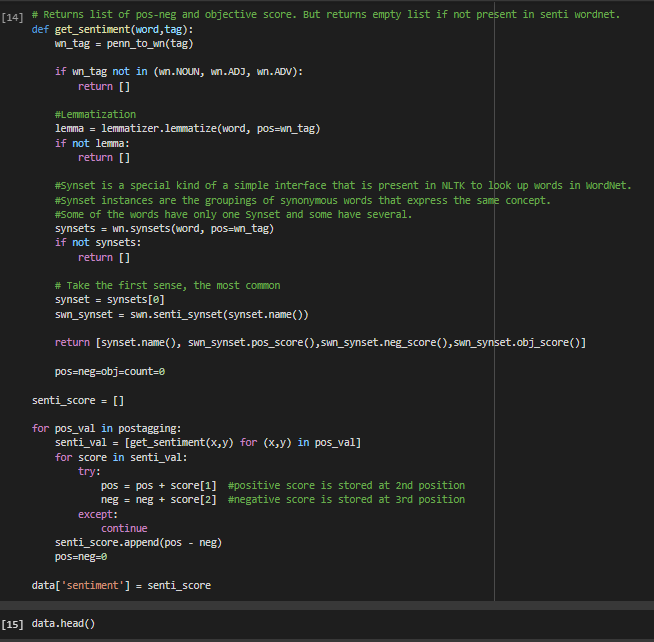
***:Lemmatization :***

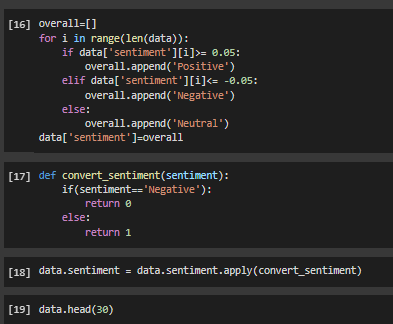
****

***: Data Preprocessing :***

***Sentiment Sentence Extraction & Pos Tagging:***

******

****

****

****

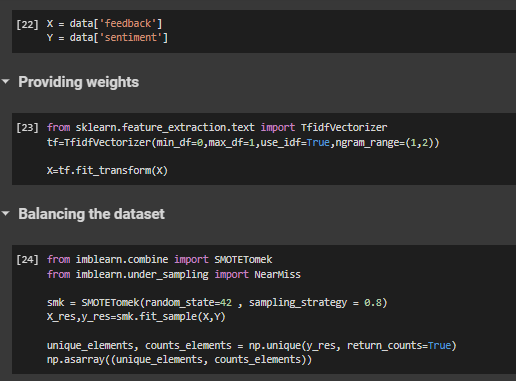
***: Sentiment Extraction :***

***Word Cloud Creation :***

****

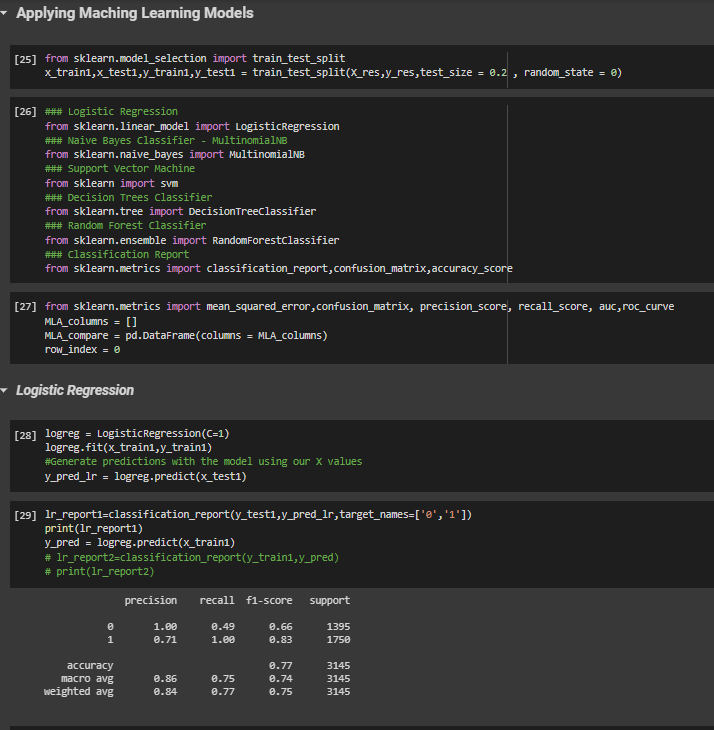
***: Word Cloud :***

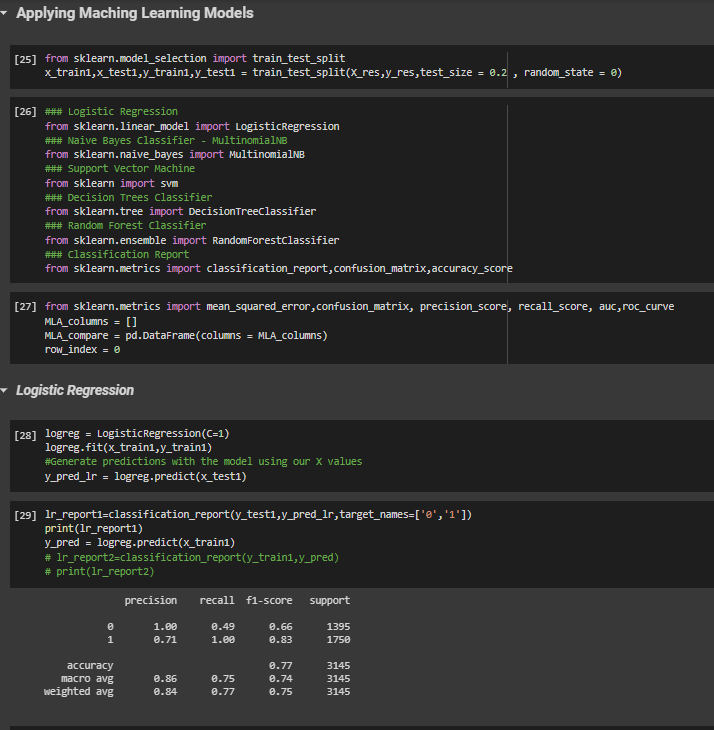
***Balancing The Dataset :***

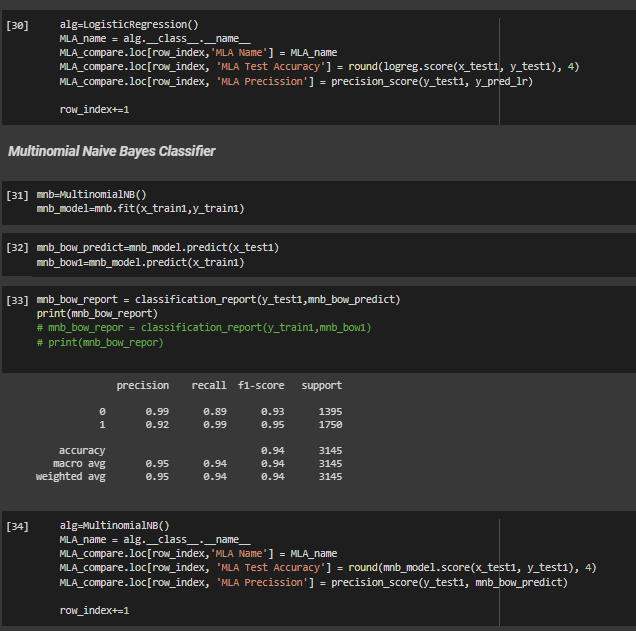
****

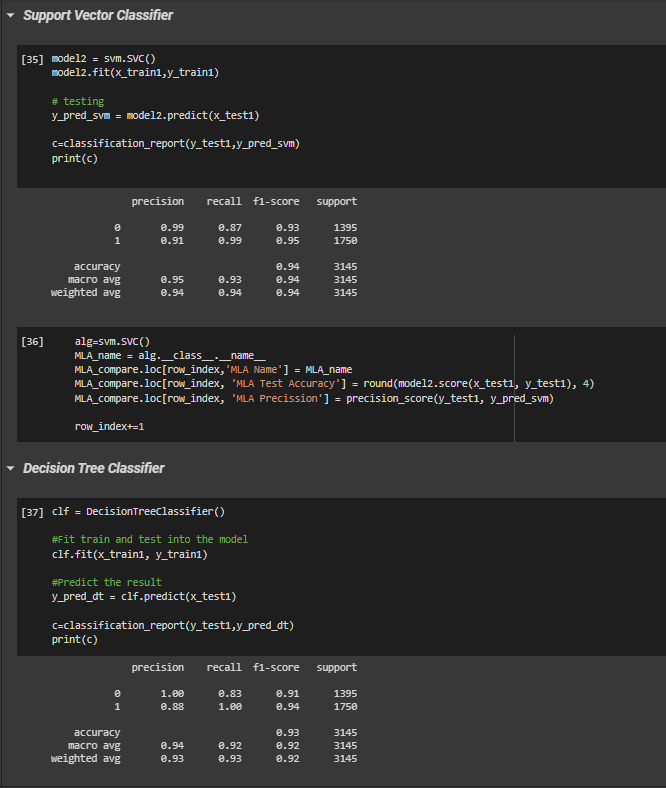
***: Balancing Dataset & Providing Weights :***

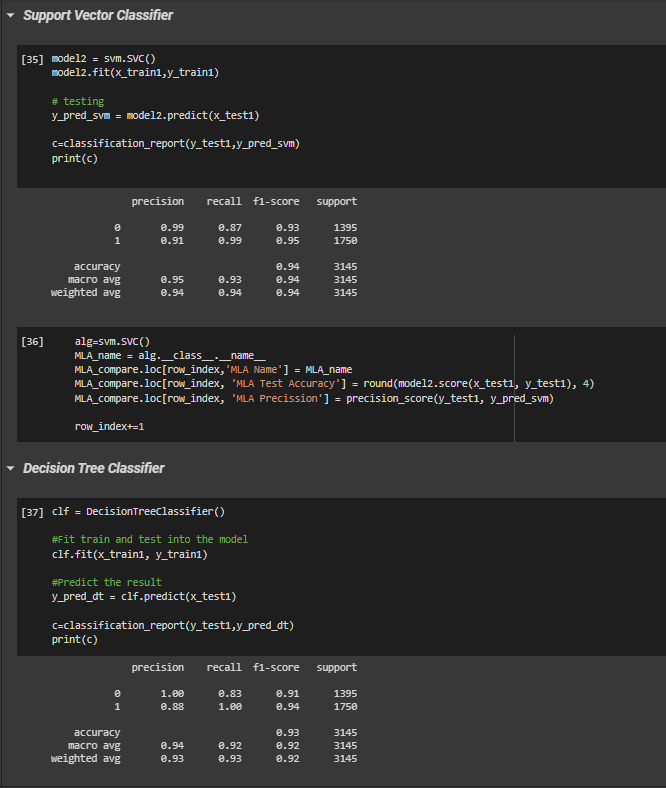
***MACHINE LEARNING Models :***

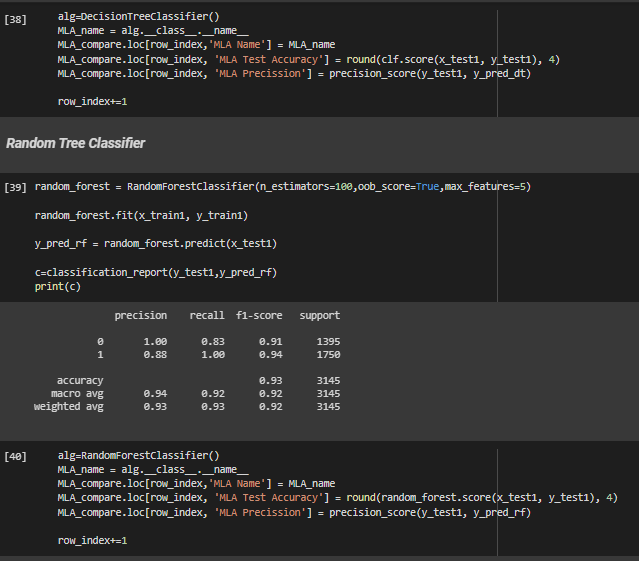
****

****

****

****

****

****

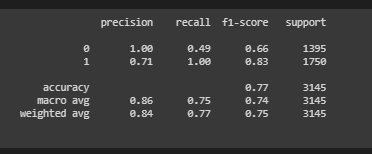
***: ML Classifiers :***

From above screenshots posted, we can say that, we have done all the precautionary steps and have applied various Machine Learning Classifiers and now we will look forward to analyze the results obtained.

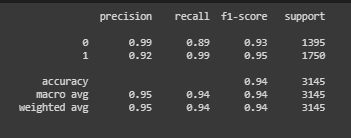
**Chapter 04: RESULTS**

**4.1 Discussion on the Results Achieved**

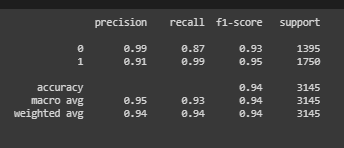
**Logistic Regression: -**



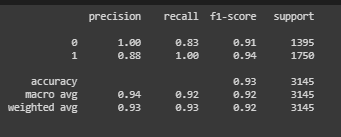
**Multinomial Naïve Bayes Classifier: -**



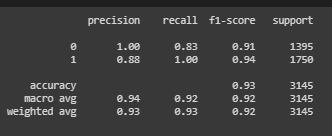
**Support Vector Classifier: -**



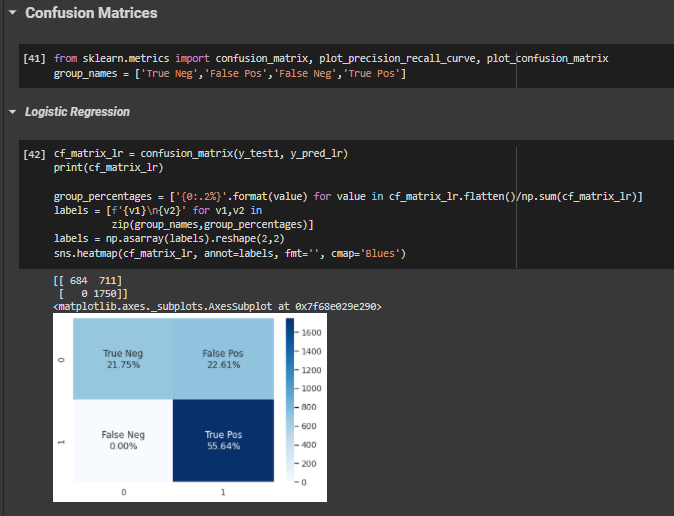
**Decision Tree Classifier: -**



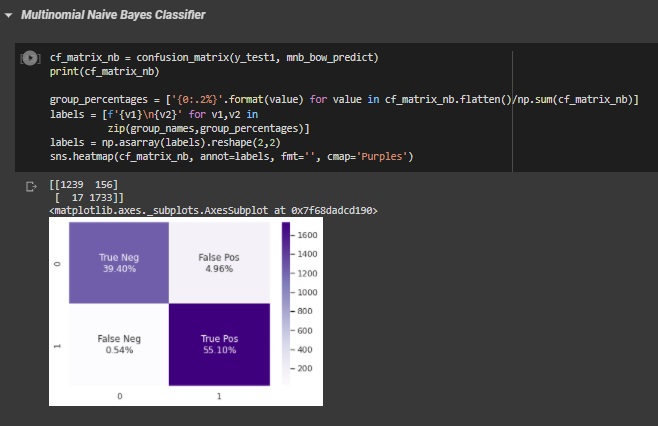
**Random Forest Classifier: -**



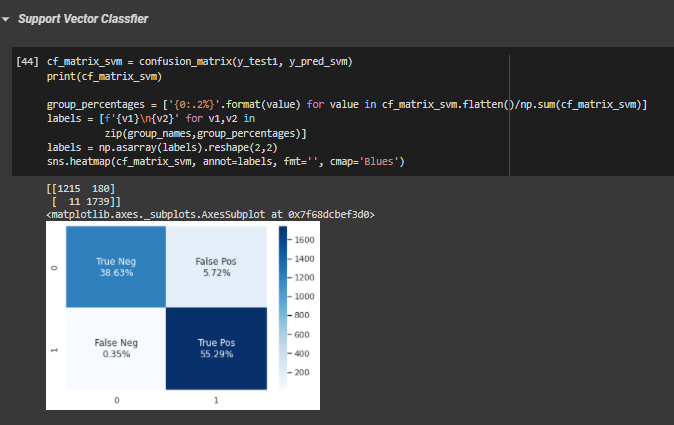
***Confusion Matrices of Classfiers :***

****

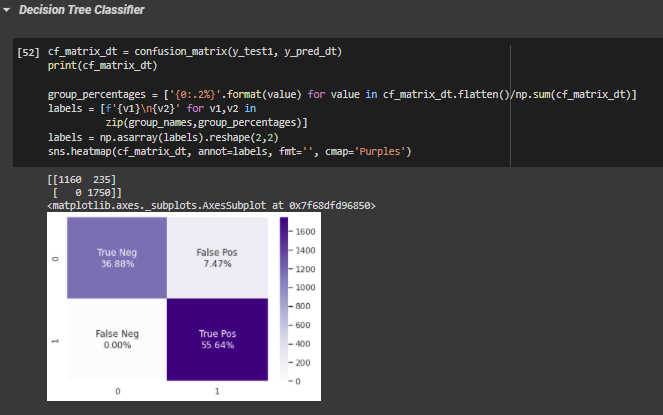
**: Logistic Regression Confusion Matrix :**

****

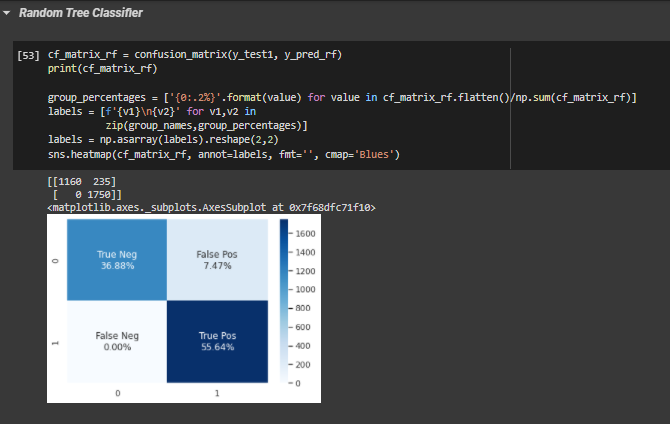
**: Multinomial Naïve Bayes Confusion Matrix :**

****

**: SVM Confusion Matrix :**

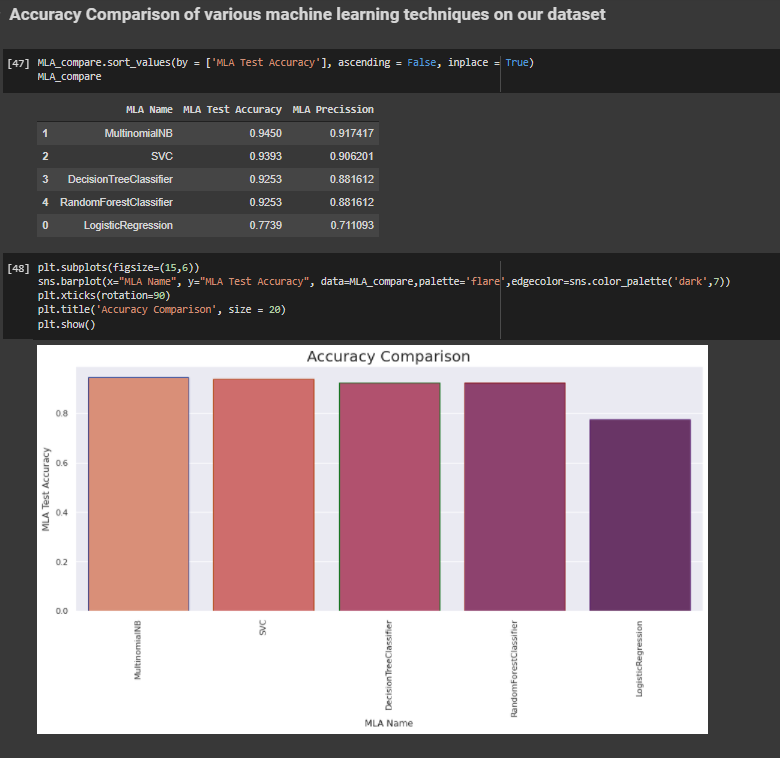
****

**: Decision Tree Confusion Matrix :**

****

**: Random Forest Confusion Matrix :**

***Accuracy Comparison of various ML techniques on our dataset:***

****

***: Accuracy Comparison :***

**RESULTS :**

After testing some arbitrary reviews, it seems that our features is performing correctly with Positive, Neutral, Negative results

We also see that after running the search, our MultinomialNB Machine Classifier has

improved to **94.50%** accuracy level

**4.2 Application of the Project**

The popular applications of our project in real life are:

Opinion mining is also a method of data investigation, which involves the

compilation, interpretation, processing and evaluation of the review to the customer.

They obtained data from various government websites in their paperwork. The collected

data is pre-processed to delete unsolicited data, and trained data are sorted by using the

Naive Bayes, Logistic Regression, Random Forest, Decision Trees and SVM algorithm.

However, these two algorithms are less accurate.

For the proposed method, the combination of Naive bays, supporting vector

machines, and ensemble algorithms are used, thereby ensuring the precision and speed of

execution of the algorithm. So, after precision measurement, if the precision is high, the

method will be used as a recommendation for the user. Then the standard evaluation so

features for the algorithms are specified. The defined data set of certain items, such as

electronics, instruments are readily read and categorized to provide good accuracy and

performance, for instance in product review data set in Amazon. To suggest an item to the

customer depends on customer feedback, the rating classification is widely used.

Sentiment analysis deals with the classification of texts based on the sentiments they

contain. This article focuses on a typical sentiment analysis model consisting of three

core steps, namely data preparation, review analysis and sentiment classification, and

describes representative techniques involved in those steps.

Sentiment analysis is an emerging research area in text mining and computational

linguistics, and has attracted considerable research attention in the past few years.

Future research shall explore sophisticated methods for opinion and product feature

extraction, as well as new classification models that can address the ordered labels

property in rating inference.

**4.3 Limitation of the Minor Project**

**4.4 Future Work**

**References**

**All references must be in APA format.( You can go to google scholar, click on the article and cite the article in APA format).**

**For example;**

1. Ravì, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., & Yang, G. Z. (2016). Deep learning for health informatics. *IEEE journal of biomedical and health informatics*, *21*(1), 4-21.

Additional Guidelines:-

1. There is no page limit as such but above mentioned titles and non-titles must be part of the project report. Ideally it should be of 25 to 50 pages.
2. Each report must be designed as per the format but the title of chapters, sub titles and number of chapters can be decided by the guide based on specific project.
3. Each project report must have a proper Table of Content.
4. Each report must have the page number starting from the chapter number 1.
5. Front page of the report should not be marked with any page number.
6. Declaration, Certificate and Table of Content must have numbers in roman (I, II, III….)
7. All references must be written in APA style.
8. All the references must be cited into the text at the appropriate place in the report.
9. Similarity index must be less than 20%