**The Project Report on**

**Sentiment Analysis for Microblogs using Machine Learning**

**Submitted To**

**Acharya Nagarjuna University**

**A project report submitted in the partial fulfillment of the requirements for the**

**Award of the Degree of**

**MASTER OF COMPUTER APPLICATIONS**

**Submitted By**

**ARUNURU NAVEEN**

**Reg.No: L20MC23014**

**Under The Esteemed Guidance Of**

**Sri N. KIRAN KUMAR M. Tech**

**Assistant Professor**

**DEPARTMENT OF MCA**

**ENGINEERING COLLEGE**

**(Approved by A.I.C.T.E)**

**(Affiliated to)**

**2019-2021**

**DEPARTMENT OF MCA**

**ENGINEERING COLLEGE**

**-522101**

**CERTIFICATE**

This is to certify that this project work entitled “**Secure Cloud Storage based on RLWE Problem”** is the bonafide work carried out by **ARUNURU NAVEEN**, **Reg.No: L20MC23014** submitted in Partial fulfillment of the requirement for the Award of Degree of “**Master of** **Computer Applications**”, during the academic year 2019-2021.

The results submitted in this project have been verified and are found to be satisfactory. The results embodied in this thesis have not been submitted to any other university for the award of the any other degree/diploma.

**Project Guide: Head of the Department:**

**Sri N. Kiran Kumar M. Tech Sri K.N. Prasad M. Tech**

**Assistant Professor Associate Professor & Head**

**Dept. of MCA Dept. of MCA**

**Bapatla Engineering College Bapatla Engineering College**

**Bapatla. Bapatla.**

**Signature of External Examiner**

# ACKNOWLEDGEMENT

I sincerely thank the following distinguished personalities who have given their advice and support for successful completion of this work.

I would like to thank **Dr. V. Damodara Naidu**, Principal, Bapatla Engineering College, Bapatla for providing the online resources and allowing me to carry out this work.

I extended my sincere thanks to **Sri K.N. Prasad**, Head of the Department for extending their cooperation and providing the required resources.

I am deeply indebted to most respect and my project guide **Sri N. Kiran Kumar**, Assistant Professor Dept. of MCA for his valuable and inspiring guidance, comments, suggestions and encouragements.

I extended my sincere thanks to all other teaching and non-teaching staff of the department who helped directly or indirectly for their cooperation and encouragement.

**ARUNURU NAVEEN**

**(L20MC23014)**

**DECLARATION**

This is to declare that the project **“Secure Cloud Storage based on RLWE Problem”** at Bapatla Engineering College has been presented by me during the academic year **2019-2021** in partial fulfillment of the requirements for the **“Master of Computer Application”**.

I also declare that this project is the result of my own efforts and that it has not been submitted to any other universities for the award of degree or diploma.

**ARUNURU NAVEEN**

**(L20MC23014)**

Table of Contents

[ACKNOWLEDGEMENT iii](#_Toc80180563)

[ABSTRACT viii](#_Toc80180564)

[CHAPTER-1 1](#_Toc80180565)

[INTRODUCTION 1](#_Toc80180566)

[1.1 INTRODUCTION 1](#_Toc80180567)

[CHAPTER- 2 5](#_Toc80180568)

[LITERATURE SURVEY 5](#_Toc80180569)

[2.1 LITERATURE REVIEW 5](#_Toc80180570)

[CHAPTER-3 7](#_Toc80180571)

[THEORETICAL BACKGROUND 7](#_Toc80180572)

[3.1 INTRODUCTION: 7](#_Toc80180573)

[3.2 INTRODUCTION TO PYTHON 12](#_Toc80180574)

[3.3 BENFITS OF PYTHON 21](#_Toc80180575)

[CHAPTER-4 30](#_Toc80180576)

[SYSTEM ANALYSIS 30](#_Toc80180577)

[4.1 EXISTING SYSTEM: 30](#_Toc80180578)

[4.1.1 DISADVANTAGES OF EXISTING SYSTEM: 30](#_Toc80180579)

[4.2 PROPOSED SYSTEM: 30](#_Toc80180580)

[4.2.1 ADVANTAGES OF PROPOSED SYSTEM: 31](#_Toc80180581)

[CHAPTER- 5 32](#_Toc80180582)

[SYSTEM DESIGN 32](#_Toc80180583)

[5.1 INTRODUCTION 32](#_Toc80180584)

[5.2 MODULES 32](#_Toc80180585)

[5.2.1 DATASET: 32](#_Toc80180586)

[5.2.2 PREPROCESSING: 32](#_Toc80180587)

[5.2.3 GRAPHS: 32](#_Toc80180588)

[5.2.4 PREDICTION: 32](#_Toc80180589)

[5.3 SYSTEM ARCHITECTURE 33](#_Toc80180590)

[5.4 UML DAIGRAMS 34](#_Toc80180591)

[5.4.1 CONSTRUCTION OF USE CASE DIAGRAMS: 37](#_Toc80180592)

[5.4.2 SEQUENCE DIAGRAMS: 40](#_Toc80180593)

[5.4.3 CLASS DIAGRAM: 42](#_Toc80180594)

[5.4.4 ACTIVITY DIAGRAM: 42](#_Toc80180595)

[CHAPTER-6 44](#_Toc80180596)

[SYSTEM REQUIREMENTS 44](#_Toc80180597)

[6.1 SYSTEM REQUIREMENTS 44](#_Toc80180598)

[6.1.1 HARDWARE REQUIREMENTS: 44](#_Toc80180599)

[6.1.2 SOFTWARE REQUIREMENTS: 44](#_Toc80180600)

[CHAPTER-7 45](#_Toc80180601)

[SYSTEM IMPLEMENTATION 45](#_Toc80180602)

[7.1 INPUT AND OUTPUT DESIGNS 45](#_Toc80180603)

[7.1.1 LOGICAL DESIGN 45](#_Toc80180604)

[7.1.2 PHYSICAL DESIGN 45](#_Toc80180605)

[7.2 INPUT & OUTPUT REPRESENTATION 46](#_Toc80180606)

[7.2.1 INPUT DESIGN 46](#_Toc80180607)

[7.2.2 OBJECTIVES 47](#_Toc80180608)

[7.2.3 OUTPUT DESIGN 47](#_Toc80180609)

[CHAPTER-8 69](#_Toc80180610)

[SYSTEM TESTING 69](#_Toc80180611)

[8.1 INTRODUCTION: 69](#_Toc80180612)

[8.2 LEVELS OF TESTING 69](#_Toc80180613)

[8.2.1 BLACK BOX TESTING 70](#_Toc80180614)

[8.2.2 WHITE BOX TESTING 72](#_Toc80180615)

[CHAPTER-9 74](#_Toc80180616)

[OUTPUT SCREENS 74](#_Toc80180617)

[CONCLUSION 75](#_Toc80180618)

[REFERENCES 76](#_Toc80180619)

**List of Figures**

**Name of the figure Pg.no**

[Figure 3. 1 Structure of cloud computing 7](#_Toc80179715)

[Figure 3. 2 Characteristics of cloud computing 9](#_Toc80179716)

[Figure 3. 3 Structure of service models 10](#_Toc80179717)

[Figure 5. 1 System Architecture 33](#_Toc80179735)

[Figure 5. 2 Use Case Diagram 39](#_Toc80179736)

[Figure 5. 3 Sequence diagram 41](#_Toc80179737)

[Figure 5. 4 Class Diagram 42](#_Toc80179738)

[Figure 5. 5 Activity Diagram 43](#_Toc80179739)

# ABSTRACT

An important point of information gathering behaviour of humans has always been to find out what other people think. With the growing availability and popularity of opinion rich resources such as online review sites and personal blogs, new opportunities and challenges arise as people now can, and do, actively use information technologies to seek out and understand the opinions of others.

Sentiment analysis deals with identifying and classifying opinions or sentiments expressed in source text. Social media is generating a vast amount of sentiment rich data in the form of tweets, status updates, blog posts etc. Sentiment analysis of this user generated data is very useful in knowing the opinion of the crowd. Twitter sentiment analysis is difficult compared to general sentiment analysis due to the presence of slang words and misspellings.

This project contributes to the sentiment analysis for classification which is helpful to analyse the information in the form of the number of tweets, where opinions are highly unstructured and sentiment are either positive, negative or neutral. For this, we first pre-process the dataset, after that extract the dataset that have some meaning. This dataset is called feature vector. Then select the feature vector list and thereafter apply the machine learning based classification algorithms namely: **Naïve Bayes** along with six other algorithms namely: “*Multinomial Naïve Bayes, Bernoulli Naïve Bayes, Logistic Regression, Stochastic Gradient Descent, Support vector clustering, Linear Support vector clustering*” which is used to evaluate the confidence level of the sentiment.

.

.

# CHAPTER-1

# INTRODUCTION

## Introduction

Behind a situation, to detect the view or emotion sentiment analytics inspect and discover the feeling and aim behind a bit of work or articulation or any manner of reporting. Humans speak different languages on this earth; language is an essential tool to express our views and feelings. Whatever humans say by language sentiment is associated with that. The sentiment can be constructive or pessimistic or impartial as well [1]. Like take an example, a restaurant business company put on sale various food items like milkshakes, burgers, sandwiches, pizza, etc. They also design a website to reach customers, right now where the customer could order any type of food according to their choice, and they can give reviews and suggestions any time, to upgrade the value of food or may they like the taste food or they hate the taste [2]. By using these examples we can have 3 types of reviews given by customers. First, the constructive one signifies that the food is liked by customers. Second, the review going to be a rejection, and hence company should focus on developing a strategy to improve the food. The third one going to be like the customer neither show any reaction, we can consider it as a neutral statement. By examining all the reviews the company should focus on the increase the quality and value of the food or promoting different strategies of brand awareness to increase their annual sales. But there are thousands and millions of reviews of their food and over time it's impossible to scan each customer's review and come to any decision making [4]. To solve this problem that time sentiment analytics comes to play, which analyzes huge amounts of reviews and helps to make the decision for future progress by using real-life evidence better than based on a tiny sample of data, and that time natural processing and machine learning (ML)help to analyze it.

**Algorithm**

**KNN Classifier Algorithm:**

K-nearest neighbor method can be used for both regression and classification predictive problems. This method helps in interpret output, calculate time and predictive power. The Machine learning techniques are used in various fields. KNN is also one of the machine learning method. This is also called as method of sample-based learning. This will contain the data of past datasets and can be used while predicting the new datasets. This will apply function called as distance function like Manhattan or Euclidean distance. This can be used to compute distance from samples to all other training samples. It calculates the target value for new samples. The target vale will be the weighted sum of target values of the k nearest neighbours. The valve of K can be directly proportional to the prediction. Whenever the valve of K is small this indicates there is high variance and there is low bias. If the valve of the K is larger than this indicates that there is low variance and high bias. The main advantage of this KNN is it does not require any training or the optimization. This KNN uses data samples when predicting the new datasets. Hence it is having higher complexity and also more time consumption.

This work represents a review of K-NN technique for the early prediction of food recommendation. K-NN analysis is used for predicting the unknown parameter from the known parameters. In this work we are considering vitamins as input parameters which are the main parameters to be considered for a good food recommendation, although there are many other factors that can be considered [17]. The unknown value of vitamis can be predicted from the nearest known values of the nearest neighbors by calculation of Euclidean distance between them. Then we would be able to predict type of food for given vitamin parameters. To measure the distance between points in a feature space, various distance functions can be used, in which the Euclidean distance function is the most widely used one[18]. Let p and q are represented as feature vectors. To calculate the distance between p and q, the Euclidean metric is generally used by if a=(a1, a2) and b=(b1,b2) then the distance is given

**Classifications Algorithms**

Onto the part you’ve probably been waiting for all this time: training machine learning algorithms. To be able to test the performance of our algorithms, I first performed an

80/20 train-test split, splitting our balanced data set into two pieces. To avoid overfitting, I used the very common resampling technique of k-fold cross-validation. This simply means that you separate your training data into k parts (folds) and then fit your model on k-1 folds before making predictions for the kth hold-out fold. You then repeat this process for every single fold and average the resulting predictions.

To get a better feeling of which algorithm would perform best on our data, let’s quickly spot-check some of the most popular classification algorithms:

* Logistic Regression
* Linear Discriminant Analysis
* K Nearest Neighbors (KNN)
* Classification Trees
* Support Vector Classifier
* Random Forest Classifier

**Logistic regression**: This is a classification function that uses class for building and uses a single multinomial logistic regression model with a single estimator. Logistic regression usually states where the boundary between the classes exists, also states the class probabilities depend on distance from the boundary, in a specific approach. This moves towards the extremes (0 and 1) more rapidly when data set is larger. These statements about probabilities which make logistic regression more than just a classifier. It makes stronger, more detailed predictions, and can be fit in a different way; but those strong predictions could be wrong. Logistic regression is an approach to prediction, like Ordinary Least Squares (OLS) regression. However, with logistic regression, prediction results in a dichotomous outcome [13]. Logistic regression is one of the most commonly used tools for applied statistics and discrete data analysis. Logistic regression is linear interpolation

**NLTK**

The Natural Language Toolkit, or more commonly NLTK, is a suite of libraries and programs for symbolic and statistical natural language processing (NLP) for the Python programming language. NLTK includes graphical demonstrations and sample data. It is accompanied by a book that explains the underlying concepts behind the language processing tasks supported by the toolkit, plus a cookbook. NLTK is intended to support research and teaching in NLP or closely related areas, including empirical linguistics, cognitive science, artificial intelligence, information retrieval, and machine learning. NLTK has been used successfully as a teaching tool, as an individual study tool, and as a platform for prototyping and building research systems.

# CHAPTER- 2

# literature survey

## 2.1 literature review

**Literature survey is the most important step in software development process. Before developing the tool it is necessary to determine the time factor, economy and team strength. Once these things are satisfied, then next steps is to determine which operating system and language can be used for developing the tool. Once the programmers start building the tool the programmers need lot of external support. This support can be obtained from senior programmers, from book or from websites. Before building the system the above consideration are taken into account for developing the proposed system. We have to analyze Machine learning and sentimental analysis: With the rise of Social Media internet users became able to easily express and share their opinions about companies, products, services, events etc. Thus companies became interested in monitoring what people say about their brands in order to get feedback or enhance their marketing efforts. Machine Learning has several interesting applications in Social Media Monitoring. It is used in order to evaluate the opinions of the users and classify them as positive, negative or neutral (Also known as Sentiment Analysis). In addition it can be used to detect whether the posts are objective or subjective, what is the natural language of the posts and whether the posts were written by a man or woman.**. **Instead of identifying the sentiment of the entire text based on the specific subject, NLP may identify the text fragment using subject and sentiment lexicons to do sentiment classification. One of the NLP approaches is the feature extraction algorithm. It may be extracted subject-specific characteristics, extracted sentiment from each sentiment-bearing lexicon, and each sentiment's relationship to a certain subject. It beat machine learning methods, with 87 percent accuracy for online review articles and 919 percent accuracy for evaluations of ordinary web pages and news items. This method concentrated on generic text and eliminated several challenging circumstances in order to get better results, such as confusing sentences or sentences with no feeling. This study proposes a method for doing Twitter sentiment analysis based on a specific topic. Many preprocessing processes were used to remove noise from tweets and display them in formal language. NLP is used to determine the sentiment of tweets by identifying the subjective element of tweets that is associated with the subject and classifying the sentiment of tweets. Positive, negative, and neutral labels will be assigned to the text. This research examines related literature released between 2014 and 2019 to gain a better knowledge of sentiment analysis' applicability in social media text platforms. Analyzing public opinion is a technique that employs Natural Language Processing (NLP) to Extract, transform, and analyses sentiment in a text to classify it as positive, negative, or natural [4]. The majority of past research used sentiment analysis to better understand**

# Chapter-3

# Theoretical background

## 3.1 Introduction:

**What is Machine Learning?**

Tom Mitchell states machine learning as “A computer program is said to learn from experience and from some tasks and some performance on, as measured by, improves with experience”. Machine Learning is combination of correlations and relationships, most machine learning algorithms in existence are concerned with finding and/or exploiting relationship between datasets. Once Machine Learning Algorithms can pinpoint on certain correlations, the model can either use these relationships to predict future observations or generalize the data to reveal interesting patterns. In Machine Learning there are various types of algorithms such as Regression, Linear Regression, Logistic Regression, Naive Bayes Classifier, Bayes theorem, KNN (K-Nearest Neighbor Classifier), Decision Tress, Entropy, ID3, SVM (Support Vector Machines), K-means Algorithm, Random Forest and etc.,

The name machine learning was coined in 1959 by Arthur Samuel. Machine learning explores the study and construction of algorithms that can learn from and make predictions on data Machine learning is closely related to (and often overlaps with) computational statistics, which also focuses on prediction-making through the use of computers. It has strong ties to mathematical optimization, which delivers methods, theory and application domains to the field. Machine learning is sometimes conflated with data mining, where the latter subfield focuses more on exploratory data analysis and is known as unsupervised learning.

Within the field of data analytics, machine learning is a method used to devise complex models and algorithms that lend themselves to prediction; in commercial use, this is known as predictive analytics. These analytical models allow researchers, data scientists, engineers, and analysts to "produce reliable, repeatable decisions and results" and uncover "hidden insights" through learning from historical relationships and trends in the data.

Machine learning tasks Machine learning tasks are typically classified into several broad categories:

Supervised learning: The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs. As special cases, the input signal can be only partially available, or restricted to special feedback.

Semi-supervised learning: The computer is given only an incomplete training signal: a training set with some (often many) of the target outputs missing.

Active learning: The computer can only obtain training labels for a limited set of instances (based on a budget), and also has to optimize its choice of objects to acquire labels for. When used interactively, these can be presented to the user for labelling.

Unsupervised learning: No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning).

Reinforcement learning: Data (in form of rewards and punishments) are given only as feedback to the program's actions in a dynamic environment, such as driving a vehicle or playing a game against an opponent.

**Benefits of Machine Learning:**

1. **Achieve economies of scale** – increase volume output or productivity with fewer people. Your cost per unit, project or product plummets.
2. **Reduce spending on technology infrastructure.** Maintain easy access to your information with minimal upfront spending. Pay as you go (weekly, quarterly or yearly), based on demand.
3. **Globalize your workforce on the cheap.** People worldwide can access the cloud, provided they have an Internet connection.
4. **Streamline processes.** Get more work done in less time with less people.
5. **Reduce capital costs.** There’s no need to spend big money on hardware, software or licensing fees.
6. **Improve accessibility.** You have access anytime, anywhere, making your life so much easier!
7. **Monitor projects more effectively.** Stay within budget and ahead of completion cycle times.
8. **Less personnel training is needed.** It takes fewer people to do more work on a cloud, with a minimal learning curve on hardware and software issues.
9. **Minimize licensing new software.** Stretch and grow without the need to buy expensive software licenses or programs.
10. **Improve flexibility.** You can change direction without serious “people” or “financial” issues at stake.

**Advantages:**

1. **Price:** Pay for only the resources used.
2. **Security**: Cloud instances are isolated in the network from other instances for improved security.
3. **Performance:** Instances can be added instantly for improved performance. Clients have access to the total resources of the Cloud’s core hardware.
4. **Scalability:** Auto-deploy cloud instances when needed.
5. **Uptime:** Uses multiple servers for maximum redundancies. In case of server failure, instances can be automatically created on another server.
6. **Control:** Able to login from any location. Server snapshot and a software library lets you deploy custom instances.
7. **Traffic:** Deals with spike in traffic with quick deployment of additional instances to handle the load.

## 3.2 Introduction to PYTHON

**Python Technology**

Python technology is both a programming language and a platform.

**The python Programming Language**

THE PYTHON PROGRAMMING LANGUAGE IS A HIGH-LEVEL LANGUAGE THAT CAN BE CHARACTERIZED BY ALL OF THE FOLLOWING BUZZWORDS:

* + - Simple
    - Architecture neutral
    - Object oriented
    - Portable
    - Distributed
    - High performance
    - Interpreted
    - Multithreaded
    - Robust
    - Dynamic
    - Secure

With most programming languages, you either compile or interpret a program so that you can run it on your computer. The Python programming language is unusual in that a program is both compiled and interpreted. With the compiler, first you translate a program into an intermediate language called Python byte codes —the platform-independent codes interpreted by the interpreter on the Python platform. The interpreter parses and runs each Python byte code instruction on the computer. Compilation happens just once; interpretation occurs each time the program is executed. The following figure illustrates how this works.

FEATURES OF MACHINE LEARNING

• It is nothing but automating the Automation.

• Getting computers to program themselves.

• Writing Software is bottleneck.

• Machine leaning models involves machines learning from data without the help of humans or any kind of human intervention.

• Machine Learning is the science of making of making the computers learn and act like humans by feeding data and information without being explicitly programmed.

• Machine Learning is totally different from traditionally programming, here data and output is given to the computer and in return it gives us the program which provides solution to the various problems. Below is the figure.

**Traditional Programming vs Machine Learning**

• Machine Learning is a combination of Algorithms, Datasets, and Programs.

• There are Many Algorithms in Machine Learning through which we will provide us the exact solution in predicting the disease of the patients.

• How Does Machine Learning Works?

• Solution to the above question is Machine learning works by taking in data, finding relationships within that data and then giving the output.

**Machine Learning Model**

• There are various applications in which machine learning is implemented such as Web search, computing biology, finance, e-commerce, space exploration, robotics, social networks, debugging and much more.

• There are 3 types of machine learning supervised, unsupervised, and reinforcement.

**BENEFITS OF PYTHON**

• Presence of Third-Party Modules

• Extensive Support Libraries

• Open Source and Community Development

• Learning Ease and Support Available

• User-friendly Data Structures

• Productivity and Speed

• Highly Extensible and Easily Readable Language.

**Python**

Python is high level language and it is also integrated version of the program. Python is an object-oriented approach and its main aim to help programmers to write the code clearly, logical code for small and large scale of project.

Pytrhon is dynamically typed and garbage collected it also support multiple programming and it is both procedure and object oriented and also functional programming. And structural programming also supported. It has many built in function it also supports filter, map and reduce function. All the machine learning algorithm and the libraries are being supported by the python programming language. Python also support list, dict, sets and other generators. Python code can be run in different platform such as anaconda, PyCharm etc.

The main goal of this programing language is as follows:

• Python is simple, object-oriented programming language.

• The language and implementation should provide support for software engineering principles such as strong type library preset for different machine learning algorithm, and all other algorithm in simple manner.

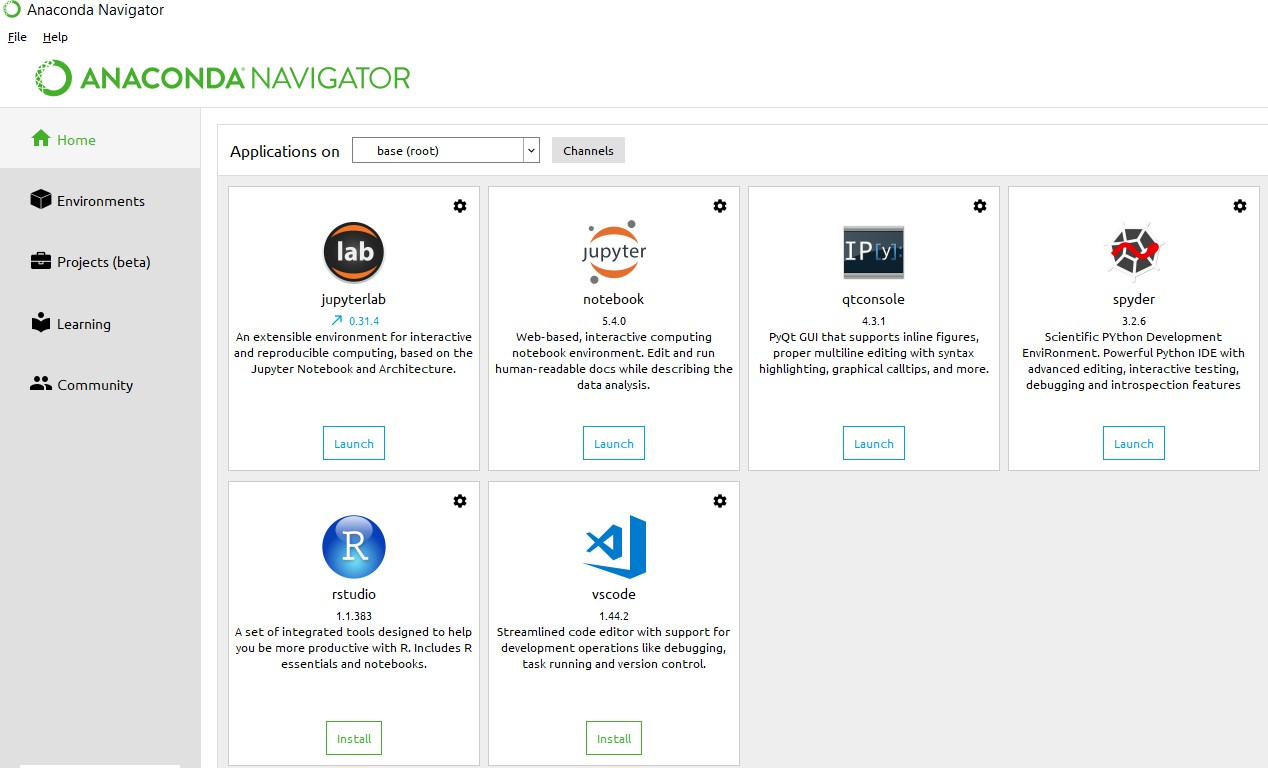
• Coding will be smooth in python and the data analysis can be easily done in python.

This is so much so to the point where we now have modules and APIs at our disposal, and you can engage in machine learning very easily without almost any knowledge at all of how it works. With the defaults from Scikit-learn, you can get 90-95% accuracy on many tasks right out of the gate. Machine learning is a lot like a car, you do not need to know much about how it works in order to get an incredible amount of utility from it.

Despite the apparent age and maturity of machine learning, I would say there's no better time than now to learn it, since you can actually use it. Machines are quite powerful, the one you are working on can probably do most of this series quickly. Data is also very plentiful lately.

**Anaconda**

Anaconda is free and open-source distribution of the Python and R programming languages for scientific computing (data science, machine Learning applications, Large- scale data processing, predictive analytics, etc.), that aims to simplify package management and deployment. It is developed and maintained by Anaconda, Inc. The distribution incudes data-science packages suitable for Windows, Linux, and macOS. Packaged versions are required and are managed by the package management system anaconda. This package manager was spun out as a separate open-source package as it ended up being useful on its own and for other things than Python. There is also a small, bootstrap version of Anaconda called Miniconda, which includes only conda, Python, the packages they depends on, and a small number of other packages.



**Anaconda Console**

**Jupyter notebook**

Jupiter Notebook or so called IPython Notebook is an interactive web based computational mean for starting with Jupiter Notebook documents. The term notebook itself is a huge entity to represent the integration with different entity sets. JSON is the main document form from the same for the execution which follows the brief on the schema and the input and output means. It has high integration with several language set and has various flexibilities with the choices. The extension used for the same is “.ipynb” which runs in this platform. It’s an open-source software package with interactive communication means. It has it’s open standards for the same. It’s an open community best for budding programmers . The flexibility of the same is phenomenon and splendidly done the configuration and integration of the same is simplest and easy on hold so that no prior distortion is generated and the efficiency of the same is measured through out any system of choice.

It’s the best software sets that been used across cross for designing and developing of the products and support wide help support. Not only to that, it provides scalability in the code and the deployment of the same. Various Language can be changed and the project can be undertaken on the same. The created notebook files can be shared and stored in various means for further utilization. It supports cultivated and interactive output sets. Easily crossed over for graphing, plotting and visualizing of the elements. Data Integration of the same is to it’s best. The integration of big data and it can process chunks of values in an approx. time which gives a better performance and the higher computational means. Various works on data like cleaning, cleansing, transforming modeling and visualizing can be done by the same

Machine learning is the ability that gives the computer to learn without being explicitly programmed. There are two types of machine learning:

Supervised Learning: supervised learning is the learning of the labelled data. It is the types of machine learning that maps the input and output based on the examples input-output pairs. In supervised learning each training data having pairs of input and desired outputs values. Supervised learning algorithm analyzes the training data and produces a function which can be used for mapping of new data.

Fig 2.1 Supervised Learning The output to solve the supervised learning algorithm are as:

• Determine the types of data, before doing anything else the user should understand which types of data set is to be used for training the data.

• Gathered the training data sets either in form of human experts or from measurements.

• Determine the feature of inputs from the learned data and depends on the inputs it changed into feature vector; number of features should not be large but should contains enough information to accurately predict the outputs.

• Check the learned function and the learned algorithm for example we use support vector machines or decisions tree.

• Complete the design and run the trained data sets.

• Analyzed the output and verify the data sets to get the accurate outputs.

Unsupervised Learning:

Unsupervised learning is a type of machine learning that helps in finding the previously unknown patterns in the data set without any known labels. It is known as self- organization and allows modelling probability densities of given inputs.

Fig 2.2 unsupervised Learning Some of the algorithm used in unsupervised learning are:

• Clustering

• Anomaly detection

• Neural networks

• Approach for learning latent variable models

• Non labelled data

Semi Supervised Machine Learning algorithm: It’s like the middle man which have some labeled data and some unlabeled which can be prosed by the both the structured and unsupervised learning.

The algorithms have been compared based upon the parameters: Size of the dataset and Number of technical indicators used. Accuracy and F-measure values have been computed for each algorithm. Long term model has been used to compute the accuracy and F-measure.

Reinforcement Learning: This type of learning is used to reinforce or strengthen the network based on critic information. That is, a network being trained under reinforcement learning, receives some feedback from the environment. However, the feedback is evaluative and not instructive as in the case of supervised learning. Based on this feedback, the network performs the adjustments of the weights to obtain better critic information in future.

This learning process is similar to supervised learning but we might have very less information. The following figure gives the block diagram of reinforcement learning:

**import numpy as np**

* NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.
* At the core of the NumPy package, is the ndarray object. This encapsulates n-dimensional arrays of homogeneous data types, with many operations being performed in compiled code for performance. There are several important differences between NumPy arrays and the standard Python sequences:
  + NumPy arrays have a fixed size at creation, unlike Python lists (which can grow dynamically). Changing the size of an ndarray will create a new array and delete the original.
  + The elements in a NumPy array are all required to be of the same data type, and thus will be the same size in memory. The exception: one can have arrays of (Python, including NumPy) objects, thereby allowing for arrays of different sized elements.
  + NumPy arrays facilitate advanced mathematical and other types of operations on large numbers of data. Typically, such operations are executed more efficiently and with less code than is possible using Python’s built-in sequences.
  + A growing plethora of scientific and mathematical Python-based packages are using NumPy arrays; though these typically support Python-sequence input, they convert such input to NumPy arrays prior to processing, and they often output NumPy arrays. In other words, in order to efficiently use much (perhaps even most) of today’s scientific/mathematical Python-based software, just knowing how to use Python’s built-in sequence types is insufficient - one also needs to know how to use NumPy arrays.

**import pandas as pd**

* Pandas is a high-level data manipulation tool developed by Wes McKinney. It is built on the Numpy package and its key data structure is called the DataFrame. DataFrames allow you to store and manipulate tabular data in rows of observations and columns of variables.
* Pandas is a high-level data manipulation tool developed by Wes McKinney. It is built on the Numpy package and its key data structure is called the DataFrame. DataFrames allow you to store and manipulate tabular data in rows of observations and columns of variables.

**import matplotlib**

* Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002.
* One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

**import seaborn as sns**

* Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with panda’s data structures.
* Seaborn helps you explore and understand your data. Its plotting functions operate on data frames and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots. Its dataset-oriented, declarative API lets you focus on what the different elements of your plots mean, rather than on the details of how to draw them

**import sklearn**

* In general, a learning problem considers a set of n samples of data and then tries to predict properties of unknown data. If each sample is more than a single number and, for instance, a multi-dimensional entry (aka multivariate data), it is said to have several attributes or features.
* Learning problems fall into a few categories:
  + supervised learning, in which the data comes with additional attributes that we want to predict (Click here to go to the scikit-learn supervised learning page).This problem can be either:
* classification: samples belong to two or more classes and we want to learn from already labeled data how to predict the class of unlabeled data. An example of a classification problem would be handwritten digit recognition, in which the aim is to assign each input vector to one of a finite number of discrete categories. Another way to think of classification is as a discrete (as opposed to continuous) form of supervised learning where one has a limited number of categories and for each of the n samples provided, one is to try to label them with the correct category or class.
* regression: if the desired output consists of one or more continuous variables, then the task is called regression. An example of a regression problem would be the prediction of the length of a salmon as a function of its age and weight.
  + unsupervised learning, in which the training data consists of a set of input vectors x without any corresponding target values. The goal in such problems may be to discover groups of similar examples within the data, where it is called clustering, or to determine the distribution of data within the input space, known as density estimation, or to project the data from a high-dimensional space down to two or three dimensions for the purpose of visualization (Click here to go to the Scikit-Learn unsupervised learning page).

**import imblearn**

* In machine learning, while building a classification model we sometimes come to situations where we do not have an equal proportion of classes. That means when we have class imbalance issues for example, we have 500 records of 0 class and only 200 records of 1 class. This is called a class imbalance. All machine learning models are designed in such a way that they should attain maximum accuracy but in these types of situations, the model gets biased towards the majority class and will, at last, reflect on precision and recall. So how to build a model on these types of data set in a manner that the model should correctly classify the respective class and does not get biased.
* To get rid of these imbalance class issues few techniques are used called as Imblearn Technique that is mainly used in these types of situations. Imblearn techniques help to either up sample the minority class or down sample the majority class to match the equal proportion. Through this article, we will discuss imblearn techniques and how we can use them to do up sampling and down sampling. For this experiment, we are using Pima Indian Diabetes data since it is an imbalance class data set. The data is available on Kaggle for downloading.

**import sklearn.metrics as m**

There are 3 different APIs for evaluating the quality of a model’s predictions:

• Estimator score method: Estimators have a score method providing a default evaluation criterion for the problem they are designed to solve. This is not discussed on this page, but in each estimator’s documentation.

• Scoring parameter: Model-evaluation tools using cross-validation (such as model\_selection.cross\_val\_score and model\_selection.GridSearchCV) rely on an internal scoring strategy. This is discussed in the section The scoring parameter: defining model evaluation rules.

• Metric functions: The metrics module implements functions assessing prediction error for specific purposes. These metrics are detailed in sections on Classification metrics, Multilabel ranking metrics, Regression metrics and Clustering metrics.

Finally, Dummy estimators are useful to get a baseline value of those metrics for random predictions.

**from sklearn.cluster import KMeans**

* Clustering of unlabelled data can be performed with the module sklearn.cluster.
* Each clustering algorithm comes in two variants: a class, that implements the fit method to learn the clusters on train data, and a function, that, given train data, returns an array of integer labels corresponding to the different clusters. For the class, the labels over the training data can be found in the labels\_ attribute.

**from imblearn.under\_sampling import RandomUnderSampler**

Imbalanced datasets are those where there is a severe skew in the class distribution, such as 1:100 or 1:1000 examples in the minority class to the majority class.

This bias in the training dataset can influence many machine learning algorithms, leading some to ignore the minority class entirely. This is a problem as it is typically the minority class on which predictions are most important.

One approach to addressing the problem of class imbalance is to randomly resample the training dataset. The two main approaches to randomly resampling an imbalanced dataset are to delete examples from the majority class, called undersampling, and to duplicate examples from the minority class, called oversampling.

In this tutorial, you will discover random oversampling and undersampling for imbalanced classification

After completing this tutorial, you will know:

• Random resampling provides a naive technique for rebalancing the class distribution for an imbalanced dataset.

• Random oversampling duplicates examples from the minority class in the training dataset and can result in overfitting for some models.

• Random undersampling deletes examples from the majority class and can result in losing information invaluable to a model.

**from sklearn.preprocessing import StandardScaler**

* Standardize features by removing the mean and scaling to unit variance
* The standard score of a sample x is calculated as:
* z = (x - u) / s
* where u is the mean of the training samples or zero if with\_mean=False, and s is the standard deviation of the training samples or one if with\_std=False.
* Centering and scaling happen independently on each feature by computing the relevant statistics on the samples in the training set. Mean and standard deviation are then stored to be used on later data using transform.
* Standardization of a dataset is a common requirement for many machine learning estimators: they might behave badly if the individual features do not more or less look like standard normally distributed data (e.g. Gaussian with 0 mean and unit variance).
* For instance many elements used in the objective function of a learning algorithm (such as the RBF kernel of Support Vector Machines or the L1 and L2 regularizers of linear models) assume that all features are centered around 0 and have variance in the same order. If a feature has a variance that is orders of magnitude larger that others, it might dominate the objective function and make the estimator unable to learn from other features correctly as expected.

**from sklearn.model\_selection import train\_test\_split**

* Before discussing train\_test\_split, you should know about Sklearn (or Scikit-learn). It is a Python library that offers various features for data processing that can be used for classification, clustering, and model selection.
* Model\_selection is a method for setting a blueprint to analyze data and then using it to measure new data. Selecting a proper model allows you to generate accurate results when making a prediction.
* To do that, you need to train your model by using a specific dataset. Then, you test the model against another dataset.
* If you have one dataset, you'll need to split it by using the Sklearn train\_test\_split function first.
* train\_test\_split is a function in Sklearn model selection for splitting data arrays into two subsets: for training data and for testing data. With this function, you don't need to divide the dataset manually.
* By default, Sklearn train\_test\_split will make random partitions for the two subsets. However, you can also specify a random state for the operation.

**import seaborn as sns**

* Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.
* For a brief introduction to the ideas behind the library, you can read the introductory notes. Visit the installation page to see how you can download the package and get started with it. You can browse the example gallery to see what you can do with seaborn, and then check out the tutorial and API reference to find out how.
* To see the code or report a bug, please visit the GitHub repository. General support questions are most at home on stackoverflow or discourse, which have dedicated channels for seaborn.

**from sklearn.model\_selection import cross\_val\_score**

* Learning the parameters of a prediction function and testing it on the same data is a methodological mistake: a model that would just repeat the labels of the samples that it has just seen would have a perfect score but would fail to predict anything useful on yet-unseen data. This situation is called overfitting. To avoid it, it is common practice when performing a (supervised) machine learning experiment to hold out part of the available data as a test set X\_test, y\_test. Note that the word “experiment” is not intended to denote academic use only, because even in commercial settings machine learning usually starts out experimentally. Here is a flowchart of typical cross validation workflow in model training. The best parameters can be determined by grid search techniques.
* **import tensorflow as tf**

To do machine learning TensorFlow, you are likely to need to define, save, and restore a model.

A model is, abstractly:

• A function that computes something on tensors (a forward pass)

• Some variables that can be updated in response to training

In this guide, you will go below the surface of Keras to see how TensorFlow models are defined. This looks at how TensorFlow collects variables and models, as well as how they are saved and restored.



**from tensorflow.keras import layers, models, optimizers, callbacks, regularizers, metrics**

When to use a Sequential model

A Sequential model is appropriate for a plain stack of layers where each layer has exactly one input tensor and one output tensor.

A Sequential model is not appropriate when:

• Your model has multiple inputs or multiple outputs

• Any of your layers has multiple inputs or multiple outputs

• You need to do layer sharing

• You want non-linear topology (e.g. a residual connection, a multi-branch model)

* **from sklearn import metrics**

Accuracy classification score.

In multilabel classification, this function computes subset accuracy: the set of labels predicted for a sample must exactly match the corresponding set of labels in y\_true.

# Chapter-4

# System analysis

## 4.1 EXISTING SYSTEM:

• . Classifying tweets into positive and negative classes using distant supervision was presented by Alec Go,Richa Bhayani and Lei Huang , 2009 .They presented an approach for automatically classifying the sentiment of twitter messages with respect to a query term . • an approach of adding semantics as additional features into training set for sentiment analysis. For each extracted entity its semantic concepts were added as an additional feature and the correlation of the representative concepts with positive / negative sentiments were calculated.

### 4.1.1 DISADVANTAGES OF EXISTING SYSTEM:

Numerous research work has been already done in field of sentiment analysis. But the informal tone of tweets has always been a challenge for the analysis. Because of the unbiased nature of tweets , twitter database has been in lime light for sentiment analysis of movies , products , popularity or anything of that sort

## 4.2 PROPOSED SYSTEM:

• In proposed system, we will attempt to conduct sentiment analysis on “tweets” using various different text preprocessing and NLP algorithm. We attempt to classify the polarity of the tweet where it is either positive or negative. If the tweet has both positive and negative elements, the more dominant sentiment should be picked as the final label.

• We use the dataset from Kaggle which was crawled and labeled positive/negative. The data provided comes with emoticons, usernames and hashtags which are required to be processed and converted into a standard form. We also need to extract useful features from the text such unigrams bigrams and trigrams which is a form of representation of the “tweet”. We use Naïve-Bayes Classifier to conduct sentiment analysis using the extracted features. However, just relying on an individual model did not give a high accuracy..

.

### 4.2.1 ADVANTAGES OF PROPOSED SYSTEM:

* Sentiment analysis also is used to monitor and analyse social phenomena, for the spotting of potentially dangerous situations and determining the general mood of the blogosphere.
* In political field, it is used to keep track of political view, to detect consistency and inconsistency between statements and actions at the government level. It can be used to predict election results as well!

# CHAPTER- 5

# SYSTEM design

## 5.1 introduction

System Design Introduction:

The System Design Document describes the system requirements, operating environment, system and subsystem architecture, files and database design, input formats, output layouts, human-machine interfaces, detailed design, processing logic, and external interfaces.

## 5.2 modules

### 5.2.1 Dataset:

* In this project se are using Twitter sentiment dataset which is taken form Kaggle website which as tweet values as features and labels as 0 and 1.

### 5.2.2 Pre-processing:

* • Features are extracted from data set and stored in variable as xtrain variable and labels are stored in y train variable. Data is preprocessing by standard scalar function and new features and labels are generated.

### 5.2.3 Methodology:

As seen from the above figure, we can see how the data is divided into different sets and then trained for different models. • The dataset was first divided into training set (80%) and pre-training set(20%). • The pre-training set was divided into pre-train(80%) and pre-test(20%) • Now, the training set is further is divided into train(80%) and validation set(20%). This train set is again divided into train(80%) and test set(20%). So, now I have train validation and test sets separate which are nonoverlapping. • The pretrain set was used to find the best models for the given dataset. I took best 4 models using pretest set. Their performance was compared based on their mean absolute errors. • Once the best 4 models were obtained, hyperparameters for these models were tuned and the best parameter was selected.

**Data preprocessing:**

* Features are extracted from data set and stored in variable as xtrain variable and labels are stored in y train variable. Data is preprocessing by standard scalar function and new features and labels are generated.

**Testing training:**

* In this stage data is sent to testing and training function and divided in to four parts x test train, and y test train. Train variables are used for passing to algorithm where as test are used for calculating accuracy of the algorithm.

**Initializing Multiple Algorithms and training with Logistic regression:**

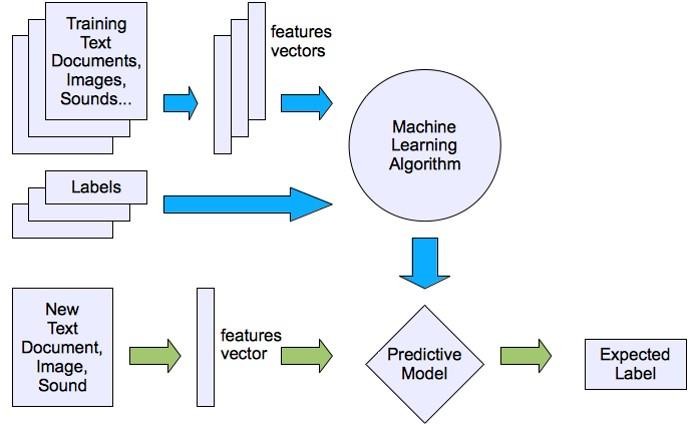
* In this stage machine learning algorithms are initialized and train values are given to algorithm by this information algorithm will know what are features and what are labels. Then data is modeled and stored as pickle file in the system which can be used for prediction.
* Data set is trained with multiple algorithms and accuracy of each model is calculated and best model is used for prediction

**Predict data:**

* In this stage new data is taken as input and trained models are loaded using pickle and then values are preprocessed and passed to predict function to find out result which is showed on web application.

## 5.3 system architecture

A system architecture or systems architecture is the conceptual model that defines the structure, behavior, and more views of a system. An architecture description is a formal description and representation of a system. Organized in a way that supports reasoning about the structures and behaviors of the system.



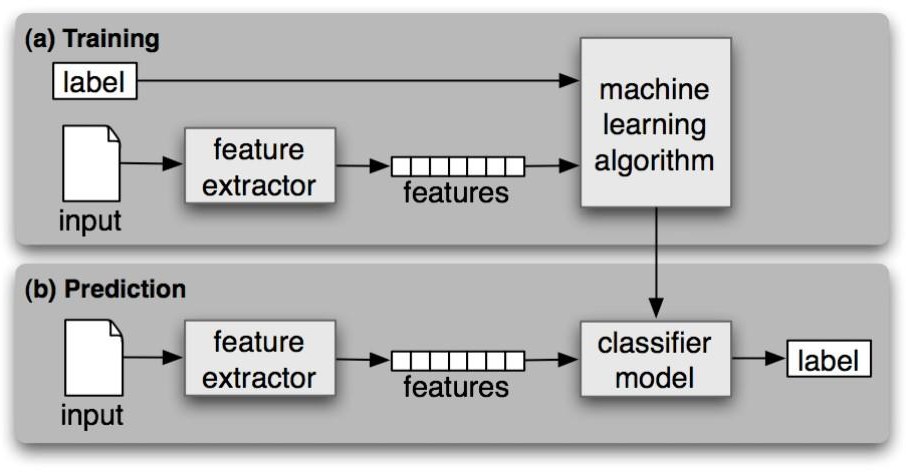


Figure 5. 1 System Architecture

3-Tier Architecture:

The three-tier software architecture (a three-layer architecture) emerged in the 1990s to overcome the limitations of the two-tier architecture. The third tier (middle tier server) is between the user interface (client) and the data management (server) components. This middle tier provides process management where business logic and rules are executed and can accommodate hundreds of users (as compared to only 100 users with the two tier architecture) by providing functions such as queuing, application execution, and database staging.

The three tier architecture is used when an effective distributed client/server design is needed that provides (when compared to the two tier) increased performance, flexibility, maintainability, reusability, and scalability, while hiding the complexity of distributed processing from the user. These characteristics have made three layer architectures a popular choice for Internet applications and net-centric information systems.

**Advantages of Three-Tier:**

* Separates functionality from presentation.
* Clear separation – better understanding.
* Changes limited to well define components.
* Can be running on WWW.
* Effective network performance.

## 5.4 UML DAIGRAMS

Global Use Case Diagrams:

Identification of actors:

Actor: Actor represents the role a user plays with respect to the system. An actor interacts with, but has no control over the use cases.

Graphical representation:



<<Actor name>>

An actor is someone or something that:

Interacts with or uses the system.

* Provides input to and receives information from the system.
* Is external to the system and has no control over the use cases.

Actors are discovered by examining:

* Who directly uses the system?
* Who is responsible for maintaining the system?
* External hardware used by the system.
* Other systems that need to interact with the system.

Questions to identify actors:

* + Who is using the system? Or, who is affected by the system? Or, which groups need help from the system to perform a task?
  + Who affects the system? Or, which user groups are needed by the system to perform its functions? These functions can be both main functions and secondary functions such as administration.
  + Which external hardware or systems (if any) use the system to perform tasks?
  + What problems does this application solve (that is, for whom)?
  + And, finally, how do users use the system (use case)? What are they doing with the system?

The actors identified in this system are:

1. System Administrator
2. Customer
3. Customer Care

Identification of use cases:

Use case: A use case can be described as a specific way of using the system from a user’s (actor’s) perspective.

Graphical representation:



A more detailed description might characterize a use case as:

* Pattern of behavior the system exhibits
* A sequence of related transactions performed by an actor and the system
* Delivering something of value to the actor

Use cases provide a means to:

* capture system requirements
* communicate with the end users and domain experts
* test the system

Use cases are best discovered by examining the actors and defining what the actor will be able to do with the system.

Guide lines for identifying use cases:

* For each actor, find the tasks and functions that the actor should be able to perform or that the system needs the actor to perform. The use case should represent a course of events that leads to clear goal
* Name the use cases.
* Describe the use cases briefly by applying terms with which the user is familiar.

This makes the description less ambiguous

Questions to identify use cases:

* What are the tasks of each actor?
* Will any actor create, store, change, remove or read information in the system?
* What use case will store, change, remove or read this information?
* Will any actor need to inform the system about sudden external changes?
* Does any actor need to inform about certain occurrences in the system?
* What usecases will support and maintains the system?

**1.2 Flow of Events**

A flow of events is a sequence of transactions (or events) performed by the system. They typically contain very detailed information, written in terms of what the system should do, not how the system accomplishes the task. Flow of events are created as separate files or documents in your favorite text editor and then attached or linked to a use case using the Files tab of a model element.

A flow of events should include:

* When and how the use case starts and ends
* Use case/actor interactions
* Data needed by the use case
* Normal sequence of events for the use case
* Alternate or exceptional flows

### 5.4.1 Construction of Use case diagrams:

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.



Figure 5. 2 Use Case Diagram

### 5.4.2 SEQUENCE DIAGRAMS:

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

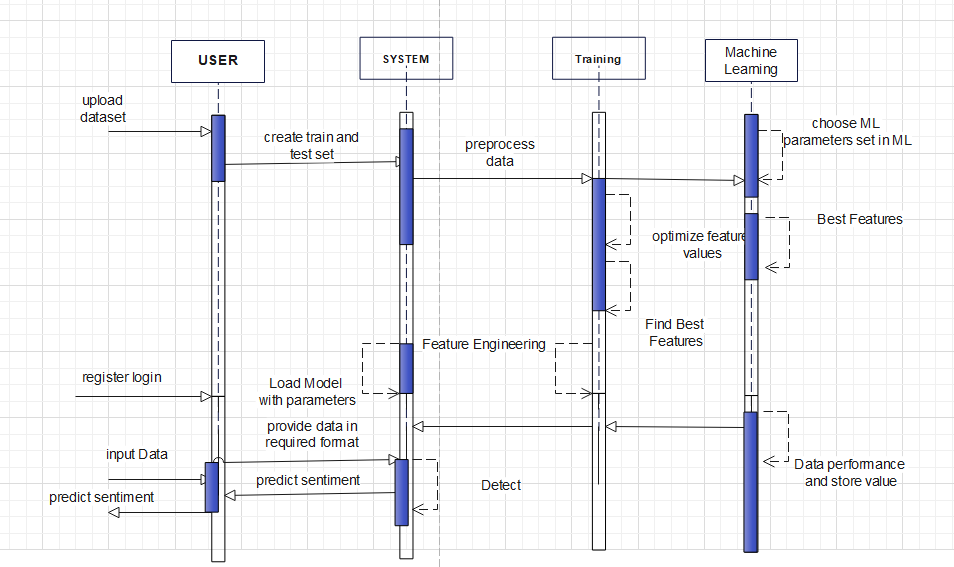


Figure 5. 3 Sequence diagram

### 5.4.3 CLASS DIAGRAM:

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

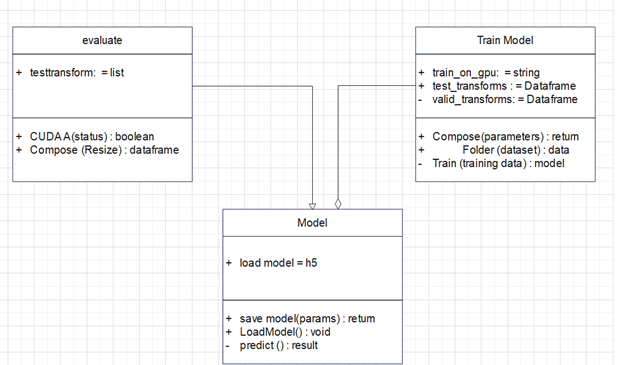


Figure 5. 4 Class Diagram

### 5.4.4 ACTIVITY DIAGRAM:

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

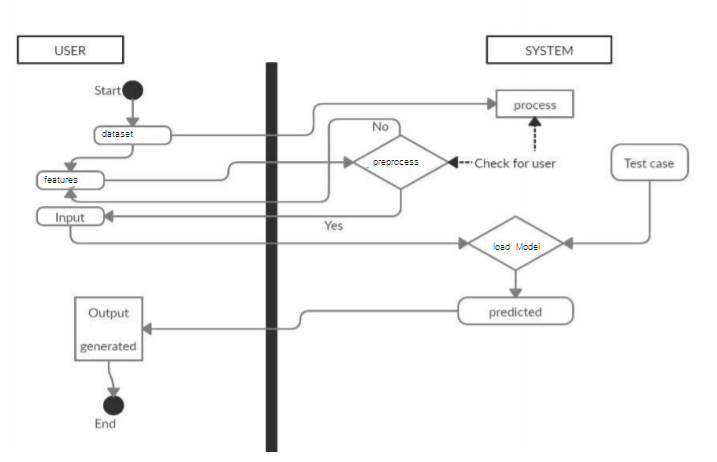
****

Figure 5. 5 Activity Diagram

REQUIREMENT ANALYSIS

FUNCTIONAL REQUIREMENTS

In software engineering, a functional requirement defines a function of a software system or its component. A function is described as a set of inputs, the behavior, and outputs. Functional requirements may be calculations, technical details, data manipulation and processing and other specific functionality that define what a system is supposed to accomplish. Behavioral requirements describing all the cases where the system uses the functional requirements are captured in use cases. The following list shows functional requirement of our project:

• We should be able to train classifier with huge amount of data in sufficient time.

• Comparison graph for training time for different algorithm is required.

• Classifier should be able to give sentiment of input given by user or from twitter.

• Fetch live tweets from twitter on mentioned topic and classify their sentiment.

• Twitter will block streaming for frequent connection and streaming. So we should be able to store sufficient tweet for each execution.

• Program should have testing option with accuracy graph as outcome of different algorithm.

• Project must have functionality to store new training data to learn more every time.

• Classifier should be trained only once to work anytime.

NON-FUNCTIONAL REQUIREMENTS

In systems engineering and requirements engineering, a non-functional requirement is a requirement that specifies criteria that can be used to judge the operation of a system, rather than specific behaviors. This should be contrasted with functional requirements that define specific behavior or functions. The plan for implementing functional requirements is detailed in the system design. The plan for implementing non-functional requirements is detailed in the system architecture.

Other terms for non-functional requirements are "constraints", "quality attributes", "quality goals", "quality of service requirements" and "non-behavioral requirements".

Some of the quality attributes are as follows:

• ACCESSIBILITY

• MAINTAINABILITY

• SCALABILITY

• PORTABILITY

ACCESSIBILITY:

Accessibility is a general term used to describe the degree to which a product, device, service, or environment is accessible by as many people as possible.

In our project people who have registered with the apps on twitter and have all four OAuth key should be able to fetch tweets from twitter using internet anytime.

Our program interface is simple and efficient and easy to use.

MAINTAINABILITY:

In software engineering, maintainability is the ease with which a software product can be modified in order to:

• Correct defects

• Meet new requirements

New functionalities can be added in the project based on the user requirements just by adding the appropriate module and functions to existing project using simple Python GUI.

Since the programming is very simple, it is easier to find and correct the defects and to make the changes in the project.

SCALABILITY:

System is capable of handling increase total throughput under an increased load of training data when resources (typically hardware) are added.

System can work normally under situations such as low bandwidth and using less hardware resources.

PORTABILITY:

Portability is one of the key concepts of high-level programming. Portability is the software code base feature to be able to reuse the existing code instead of creating new code when moving software from an environment to another.

Project can be executed under different operation conditions provided it meet its minimum configurations. Only python environment and common libraries would have to be configured in such case.

# CHAPTER-6

# system requirements

## 6.1 SYSTEM REQUIREMENTS

### 6.1.1 HARDWARE REQUIREMENTS:

* System : Intel(R) Core(TM) i3-7020U CPU @ 2.30GHz
* Hard Disk : 1 TB.
* Input Devices : Keyboard, Mouse
* Ram : 4 GB.

### 6.1.2 SOFTWARE REQUIREMENTS:

* Operating system : Windows XP/7/10.
* Coding Language : Python
* Tool : Anaconda
* Interface : Jupiter notebook

# Chapter-7

# System implementation

To conduct studies and analyses of an operational and technological nature, and To promote the exchange and development of methods and tools for operational analysis as applied to defense problems.

## 7.1 input and output designs

### 7.1.1 Logical design

The logical design of a system pertains to an abstract representation of the data flows, inputs and outputs of the system. This is often conducted via modeling, using an over-abstract (and sometimes graphical) model of the actual system. In the context of systems design are included. Logical design includes ER Diagrams i.e. Entity Relationship Diagrams

### 7.1.2 Physical design

The physical design relates to the actual input and output processes of the system. This is laid down in terms of how data is input into a system, how it is verified / authenticated, how it is processed, and how it is displayed as output. In Physical design, following requirements about the system are decided.

1. Input requirement,
2. Output requirements,
3. Storage requirements,
4. Processing Requirements,
5. System control and backup or recovery.

Put another way, the physical portion of systems design can generally be broken down into three sub-tasks:

1. User Interface Design
2. Data Design
3. Process Design

User Interface Design is concerned with how users add information to the system and with how the system presents information back to them. Data Design is concerned with how the data is represented and stored within the system. Finally, Process Design is concerned with how data moves through the system, and with how and where it is validated, secured and/or transformed as it flows into, through and out of the system. At the end of the systems design phase, documentation describing the three sub-tasks is produced and made available for use in the next phase.

Physical design, in this context, does not refer to the tangible physical design of an information system. To use an analogy, a personal computer's physical design involves input via a keyboard, processing within the CPU, and output via a monitor, printer, etc. It would not concern the actual layout of the tangible hardware, which for a PC would be a monitor, CPU, motherboard, hard drive, modems, video/graphics cards, USB slots, etc. It involves a detailed design of a user and a product database structure processor and a control processor. The H/S personal specification is developed for the proposed system.

## 7.2 INPUT & OUTPUT REPRESENTATION

### 7.2.1 Input Design

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

* What data should be given as input?
* How the data should be arranged or coded?
* The dialog to guide the operating personnel in providing input.
* Methods for preparing input validations and steps to follow when error occur.

### 7.2.2 Objectives

Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow

### Output Design

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system’s relationship to help user decision-making.

* 1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.
  2. Select methods for presenting information.
  3. Create document, report, or other formats that contain information produced by the system.

**Code**

import os

import MySQLdb

import smtplib

import random

import string

from datetime import datetime

from flask import Flask, session, url\_for, redirect, render\_template, request, abort, flash, send\_file

from database import db\_connect,inc\_reg,ins\_loginact

import pandas as pd

import io

import json

import pickle

from omicornalgo import\*

from flask import Flask, request, render\_template

import speech\_recognition as sr

from pydub import AudioSegment

from flask import Flask, render\_template, request, redirect, url\_for

import pandas as pd

from nltk.sentiment import SentimentIntensityAnalyzer

# def db\_connect():

# \_conn = MySQLdb.connect(host="localhost", user="root",

# passwd="root", db="assigndb")

# c = \_conn.cursor()

# return c, \_conn

app = Flask(\_\_name\_\_)

app.secret\_key = os.urandom(24)

## Configure upload location for audio

app.config['UPLOAD\_FOLDER'] = "./audio"

@app.route("/")

def FUN\_root():

return render\_template("index.html")

@app.route("/at.html")

def at():

return render\_template("at.html")

@app.route("/ua.html")

def ua():

return render\_template("ua.html")

@app.route("/user.html")

def ins():

return render\_template("user.html")

@app.route("/increg.html")

def increg():

return render\_template("increg.html")

@app.route("/ihome.html")

def ihome():

return render\_template("ihome.html")

# Load NLTK's SentimentIntensityAnalyzer

sia = SentimentIntensityAnalyzer()

def analyze\_sentiment(tweet):

# Perform sentiment analysis using SentimentIntensityAnalyzer

sentiment\_score = sia.polarity\_scores(tweet)

# Classify sentiment based on compound score

if sentiment\_score['compound'] >= 0.05:

return 'Positive'

elif sentiment\_score['compound'] <= -0.05:

return 'Negative'

else:

return 'Neutral'

def process\_and\_analyze\_csv(input\_file\_path, output\_file\_path):

# Load CSV dataset

df = pd.read\_csv(input\_file\_path)

# Analyze sentiment for each tweet

df['Sentiment'] = df['Tweet'].apply(analyze\_sentiment)

# Write the predictions to a new CSV file

df.to\_csv(output\_file\_path, index=False)

# Calculate total positive, negative, and neutral predictions

total\_positive = df[df['Sentiment'] == 'Positive'].shape[0]

total\_negative = df[df['Sentiment'] == 'Negative'].shape[0]

total\_neutral = df[df['Sentiment'] == 'Neutral'].shape[0]

return total\_positive, total\_negative, total\_neutral

@app.route('/analyze', methods=['POST'])

def analyze():

if request.method == 'POST':

input\_file\_path = request.form['input\_file\_path']

output\_file\_path = request.form['output\_file\_path']

# Process and analyze CSV

total\_positive, total\_negative, total\_neutral = process\_and\_analyze\_csv(input\_file\_path, output\_file\_path)

return render\_template('result.html',

total\_positive=total\_positive,

total\_negative=total\_negative,

total\_neutral=total\_neutral)

@app.route("/profile.html")

def profile():

return render\_template("profile.html")

@app.route("/index")

def index():

return render\_template("index.html")

# -------------------------------Registration-----------------------------------------------------------------

@app.route("/inceregact", methods = ['GET','POST'])

def inceregact():

if request.method == 'POST':

status = inc\_reg(request.form['username'],request.form['password'],request.form['email'],request.form['mobile'])

if status == 1:

return render\_template("user.html",m1="sucess")

else:

return render\_template("increg.html",m1="failed")

@app.route("/atact", methods = ['GET','POST'])

def atact():

if request.method == 'POST':

username = session['username']

data = request.form['data']

#predict code

col\_names = ['text']

var = pd.DataFrame(columns = col\_names)

var.loc[len(var)] = [data]

fileName="input\inputdata.csv"

var.to\_csv(fileName,index=False)

result=predict(data)

pred = result

return render\_template("at.html",m1="sucess",pred = pred, data = data )

@app.route("/uploadcsv", methods = ['GET','POST'])

def uploadcsv():

if request.method == 'POST':

if 'file' not in request.files:

return "No file part"

file = request.files['file']

if file.filename == '':

return "No selected file"

if file:

# Save the uploaded file

file.save(file.filename)

# Read the CSV file using pandas

df = pd.read\_csv(file.filename, encoding='unicode\_escape')

# Process the data as needed

# For example, you can print the first 5 rows

print(df.head())

return render\_template("at.html",m1="sucess" )

@app.route('/uaact', methods = ['GET','POST'])

def upload():

if not os.path.isdir("./audio"):

os.mkdir("audio")

if request.method == 'POST':

file = request.files['file']

fname = file.filename

file.save(os.path.join(app.config["UPLOAD\_FOLDER"], fname))

import speech\_recognition as sr

r = sr.Recognizer()

wav\_file\_pre = os.listdir("./audio")[0]

wav\_file\_pre = f"{os.getcwd()}\\audio\\{wav\_file\_pre}"

wav\_file = convert(wav\_file\_pre)

os.remove(wav\_file\_pre)

print(wav\_file)

hellow=sr.AudioFile(wav\_file)

with hellow as source:

audio = r.record(source)

try:

s = r.recognize\_google(audio)

print("Text: "+s)

#prediction code

col\_names = ['text']

var = pd.DataFrame(columns = col\_names)

var.loc[len(var)] = [s]

fileName="input\inputdata.csv"

var.to\_csv(fileName,index=False)

result=predict(s)

pred = result

except Exception as e:

print("Exception: "+str(e))

return render\_template('ua.html',text = s,pred = pred)

def convert(audio\_path):

"""

This function will convert any .wav files into .wav files compatible

with the model. -ac audio channels 1 (monochannel), -ar audio frequency 16000hz

audio\_path (str): the path associated with the .wav file that will be converted

"""

if not os.path.exists(audio\_path):

return "File Doesn't Exist"

file\_split\_list = audio\_path.split("/")

filename = file\_split\_list[-1].split(".")[0]

new\_filename = f"{filename}\_converted.wav"

file\_split\_list[-1] = new\_filename

seperator = "/"

target\_path = seperator.join(file\_split\_list)

if not audio\_path.endswith(".wav"):

return "Invalid File: Must be in .wav format"

else:

try:

os.system(f"ffmpeg -i {audio\_path} -ac 1 -ar 16000 {target\_path}")

except Exception as e:

print(e)

return

return target\_path

# #-------------------------------ADD\_END---------------------------------------------------------------------------

# # -------------------------------Loginact-----------------------------------------------------------------

@app.route("/inslogin", methods=['GET', 'POST'])

def inslogin():

if request.method == 'POST':

status = ins\_loginact(request.form['username'], request.form['password'])

print(status)

if status == 1:

session['username'] = request.form['username']

return render\_template("ihome.html", m1="sucess")

else:

return render\_template("user.html", m1="Login Failed")

# # -------------------------------Loginact End-----------------------------------------------------------------

if \_\_name\_\_ == "\_\_main\_\_":

app.run(debug=True, host='127.0.0.1', port=5000)

# Chapter-8

# System testing

## 8.1 INTRODUCTION:

Testing is the debugging program is one of the most critical aspects of the computer programming triggers, without programming that works, the system would never produce an output of which it was designed. Testing is best performed when user development is asked to assist in identifying all errors and bugs. The sample data are used for testing. It is not quantity but quality of the data used the matters of testing. Testing is aimed at ensuring that the system was accurately an efficiently before live operation commands.

Testing objectives:

The main objective of testing is to uncover a host of errors, systematically and with minimum effort and time. Stating formally, we can say, testing is a process of executing a program with intent of finding an error.

1. A successful test is one that uncovers an as yet undiscovered error.
2. A good test case is one that has probability of finding an error, if it exists.
3. The test is inadequate to detect possibly present errors.
4. The software more or less confirms to the quality and reliable standards.

## 8.2 Levels of Testing

Code testing:

This examines the logic of the program. For example, the logic for updating various sample data and with the sample files and directories were tested and verified.

Specification Testing:

Executing this specification starting what the program should do and how it should performed under various conditions. Test cases for various situation and combination of conditions in all the modules are tested.

Unit testing:

In the unit testing we test each module individually and integrate with the overall system. Unit testing focuses verification efforts on the smallest unit of software design in the module. This is also known as module testing. The module of the system is tested separately. This testing is carried out during programming stage itself. In the testing step each module is found to work satisfactorily as regard to expected output from the module. There are some validation checks for fields also. For example the validation check is done for varying the user input given by the user which validity of the data entered. It is very easy to find error debut the system.

Each Module can be tested using the following two Strategies:

1. Black Box Testing
2. White Box Testing

### 8.2.1 BLACK BOX TESTING

What is Black Box Testing?

Black box testing is a software testing techniques in which functionality of the software under test (SUT) is tested without looking at the internal code structure, implementation details and knowledge of internal paths of the software. This type of testing is based entirely on the software requirements and specifications.

In Black Box Testing we just focus on inputs and output of the software system without bothering about internal knowledge of the software program.



The above Black Box can be any software system you want to test. For example : an operating system like Windows, a website like Google ,a database like Oracle or even your own custom application. Under Black Box Testing , you can test these applications by just focusing on the inputs and outputs without knowing their internal code implementation.

Black box testing - Steps

Here are the generic steps followed to carry out any type of Black Box Testing.

* Initially requirements and specifications of the system are examined.
* Tester chooses valid inputs (positive test scenario) to check whether SUT processes them correctly. Also some invalid inputs (negative test scenario) are chosen to verify that the SUT is able to detect them.
* Tester determines expected outputs for all those inputs.
* Software tester constructs test cases with the selected inputs.
* The test cases are executed.
* Software tester compares the actual outputs with the expected outputs.
* Defects if any are fixed and re-tested.

Types of Black Box Testing

There are many types of Black Box Testing but following are the prominent ones -

* Functional testing – This black box testing type is related to functional requirements of a system; it is done by software testers.
* Non-functional testing – This type of black box testing is not related to testing of a specific functionality, but non-functional requirements  such as performance, scalability, usability.
* Regression testing – Regression testing is done  after code fixes , upgrades or any other system maintenance to check the new code has not affected the existing code.

### 8.2.2 WHITE BOX TESTING

White Box Testing is the testing of a software solution's internal coding and infrastructure. It focuses primarily on strengthening security, the flow of inputs and outputs through the application, and improving design and usability.White box testing is also known as clear, open, structural, and glass box testing.

It is one of two parts of the "box testing" approach of software testing. Its counter-part, blackbox testing, involves testing from an external or end-user type perspective. On the other hand, Whitebox testing is based on the inner workings of an application and revolves around internal testing. The term "whitebox" was used because of the see-through box concept. The clear box or whitebox name symbolizes the ability to see through the software's outer shell (or "box") into its inner workings. Likewise, the "black box" in "black box testing" symbolizes not being able to see the inner workings of the software so that only the end-user experience can be tested

WHAT DO YOU VERIFY IN WHITE BOX TESTING?

White box testing involves the testing of the software code for the following:

* Internal security holes
* Broken or poorly structured paths in the coding processes
* The flow of specific inputs through the code
* Expected output
* The functionality of conditional loops
* Testing of each statement, object and function on an individual basis

The testing can be done at system, integration and unit levels of software development. One of the basic goals of whitebox testing is to verify a working flow for an application. It involves testing a series of predefined inputs against expected or desired outputs so that when a specific input does not result in the expected output, you have encountered a bug.

HOW DO YOU PERFORM WHITE BOX TESTING?

  To give you a simplified explanation of white box testing, we have divided it into **two basic steps**. This is what testers do when testing an application using the white box testing technique:

**STEP 1) UNDERSTAND THE SOURCE CODE**

The first thing a tester will often do is learn and understand the source code of the application. Since white box testing involves the testing of the inner workings of an application, the tester must be very knowledgeable in the programming languages used in the applications they are testing. Also, the testing person must be highly aware of secure coding practices. Security is often one of the primary objectives of testing software. The tester should be able to find security issues and prevent attacks from hackers and naive users who might inject malicious code into the application either knowingly or unknowingly.

**Step 2) CREATE TEST CASES AND EXECUTE**

The second basic step to white box testing involves testing the application’s source code for proper flow and structure. One way is by writing more code to test the application’s source code. The tester will develop little tests for each process or series of processes in the application. This  method requires that the tester must have intimate knowledge of the code and is often done by the developer. Other methods include manual testing, trial and error testing and the use of testing tools as we will explain further on in this article.

Unit testing:

|  |  |
| --- | --- |
| Sl # Test Case : ­ | UTC­1 |
| Name of Test: ­ | Data Set Feature Label Extraction |
| Items being tested: ­ | Features and Labels are extracted or not |
| Sample Input: ­ | Csv files tweet dataset |
| Expected output: ­ | Features copied to x labels to y |
| Actual output: ­ | Data with features and labels are displayed. |
| **Remarks: ­** | **Pass.** |

|  |  |
| --- | --- |
| Sl # Test Case : ­ | UTC­2 |
| Name of Test: ­ | Register login |
| Items being tested: ­ | User registration details stored in Database and login success or Fail |
| Sample Input: ­ | Username, password, email, phone number, address |
| Expected output: ­ | Registration successful, login success |
| Actual output: ­ | Details stored in database and verification is success for correct details |
| Remarks: ­ | pass |

**Integration Testing:**

Integration testing is a level of software testing where individual units are combined and tested as a group. The purpose of this level of testing is to expose faults in the interaction between integrated units. Test drivers and test stubs are used to assist in Integration Testing. Integration testing is defined as the testing of combined parts of an application to determine if they function correctly. It occurs after unit testing and before validation testing. Integration testing can be done in two ways: Bottom­up integration testing and Top­down integration testing.

* + 1. **Bottom­up Integration**

This testing begins with unit testing, followed by tests of progressively higher­level combinations of units called modules or builds.

* + 1. **Top­down Integration**

In this testing, the highest­level modules are tested first and progressively, lower­level modules are tested thereafter.

In a comprehensive software development environment, bottom­up testing is usually done first, followed by top­down testing. The process concludes with multiple tests of the complete application, preferably in scenarios designed to mimic actual situations. Table 6.5 shows the test cases for integration testing and their results

|  |  |
| --- | --- |
| Sl # Test Case : ­ | ITC­1 |
| Name of Test: ­ | Input tweet text, audio ,csv file values |
| Item being tested: ­ | sentimet is predicted or not |
| Sample Input: ­ | Text audio csv |
| Expected output: ­ | Tweet sentiment which one is positive, negative or normal. |
| Actual output: ­ | tweet with sentiment values are displayed. |
| Remarks: ­ | Pass. |

|  |  |
| --- | --- |
| Sl # Test Case : ­ | ITC­2 |
| Name of Test: ­ | Predict sentiment or not |
| Item being tested: ­ | Prediction is success or not for sentiment values |
| Sample Input: ­ | sentiment predicted values |
| Expected output: ­ | sentiment with values |
| Actual output: ­ | sentiment data |
| Remarks: ­ | Pass. |

**System testing**:

System testing of software or hardware is testing conducted on a complete, integrated system to evaluate the system's compliance with its specified requirements. System testing falls within the scope of black­box testing, and as such, should require no knowledge of the inner design of the code or logic. System testing is important because of the following reasons:

System testing is the first step in the Software Development Life Cycle, where the application is tested as a whole.

The application is tested thoroughly to verify that it meets the functional and technical specifications.

The application is tested in an environment that is very close to the production environment where the application will be deployed.

System testing enables us to test, verify, and validate both the business requirements as well as the application architecture.

System Testing is shown in below tables

|  |  |
| --- | --- |
| Sl # Test Case : ­ | STC­1 |
| Name of Test: ­ | System testing in various versions of OS |
| Item being tested: ­ | OS compatibility. |
| Sample Input: ­ | Execute the program in windows XP/ Windows­7/8 |
| Expected output: ­ | Performance is better in windows­7 |
| Actual output: ­ | Same as expected output, performance is better in windows­7 |
| Remarks: ­ | Pass |

# CHAPTER-9

# Output Screens

## 9.1 Main Page:

# CONCLUSION

There is a lot of research on measuring sentiment from text because Twitter, kaggle.com, and other web portals are popular social media sites Our suggested system, which employs NLP techniques to extract subject from text and determine text polarity by evaluating sentiment lexicons related with subject, is given. The suggested system conducts Emotion Detection Algorithm based on the trials. Future study will concentrate on ways to increase sentiment analysis accuracy. It's also worth noting that, because of the prevalence of misspelt words and slang in texts and tweets, extracting sentiment can be difficult. Lexicons is difficult if the data is not preprocessed to formal language. Tweets should be converted to formal sentences in pre-processing, which is still inefficient because additional training data is required.

# REFERENCES

* 1] D. M. Muoio and C. B. Newgard, ‘‘Molecular and metabolic mechanisms of insulin resistance and β-cell failure in type 2 diabetes,’’ Nature Rev. Mol. Cell Biol., vol. 9, no. 3, pp. 193–205, Mar. 2008.
* [2] M. Maniruzzaman, M. J. Rahman, B. Ahammed, and M. M. Abedin, ‘‘Classification and prediction of diabetes disease using machine learning paradigm,’’ Health Inf. Sci. Syst., vol. 8, no. 1, pp. 1–14, Dec. 2020.
* [3] J. C. Prather, D. F. Lobach, L. K. Goodwin, J. W. Hales, M. L. Hage, and W. E. Hammond, ‘‘Medical data mining: Knowledge discovery in a clinical data warehouse,’’ in Proc. AMLA Annu. Fall Symp. Bethesda, MD, USA: American Medical Informatics Association, 1997, p. 101.
* [4] C. Azad and V. K. Jha, ‘‘Fuzzy min–max neural network and particle swarm optimization based intrusion detection system,’’ Microsyst. Technol., vol. 23, no. 4, pp. 907–918, Apr. 2017.

[5] Shivani Singh, Sonal Bait, Jayashree Rathod, Prof. Nileema Pathak,” Diabetes Prediction Using Random Forest Classifier And Intelligent Dietician ” , International Research Journal of Engineering and Technology (IRJET), Volume: 07 Issue: 01 | Jan 2020