**FINAL YEAR PROJECT REPORT**

**TERRORISM DETECTION FROM SOCIAL MEDIA**

**TRACK – 3 (INDUSTRY)**

Submitted in Partial Fulfillment for the Award of Degree of Bachelor of Technology in Computer Science and EngineeringfromRajasthan Technical University, Kota

****



**ANAND INTERNATION COLLEGE OF**

**ENGINEERING**

NEAR KANOTA, AGRA ROAD, Jaipur – 302017

**CERTIFICATE**

This is to certify that Final Year Project Report (Project Stage – I) entitled

“**TERRORISM DETECTION FROM SOCIAL MEDIA**” has been duly submitted by

**ANUJ PAREEK (16CS005)**

**JINY JAIN (16CS019)**

**MANISH RAWAT (16CS023)**

**VISHWAROOP SHAH (16CS047)**

for partial fulfillment of the Degree of Bachelor of Technology of Rajasthan Technical University.It has been found satisfactory and hence approved for submission as Project during academic session 2019-2020.

Date: 07-12-2019

**GUIDE: HEAD OF SCHOOL:**

**Prof. Anubhav Saxena Prof. Neeraj Prakash Shrivastava**

(Dept. of Computer Science & Engineering) (Dept. of Computer Science & Engineering)



**ANAND INTERNATION COLLEGE OF ENGINEERING**

NEAR KANOTA, AGRA ROAD, Jaipur – 302017

**ABSTRACT**

The content published on social media can spread vast unrest and agitation in public of any city, state or nation. In the recent years, it has been observed that before every terrorist activity, Communication /threats over the internet were found. A mechanism is needed that could judge the content that is being published. This mechanism would be useful for filtering out content that may affect the society negatively. Also, this mechanism will report us about any inappropriate conversation being held about the nation. This project is a software that will analyse the content/text/posts which being published and will inform the administrator if the content is analysed to be suspicious. This project will contribute in the well being of the society by preventing inappropriate data from being published on the internet.The basic concept is to analyze the content published using machine learning and makeit well filtered for the well being of the society.



**ANAND INTERNATION COLLEGE OF ENGINEERING**

NEAR KANOTA, AGRA ROAD, Jaipur – 302017

**DECLARATION**

We hereby declare that the report of the project entitled **TERRORISM DETECTION FROM SOCIAL MEDIA** is a record of an original work done by us at Anand International College of Engineering, **Jaipur** under the guidance of mentorship of **Prof. Anubhav Saxena** (Dept. of Computer Science and Engineering) and coordination of **Mr. Neeraj Prakash Shrivastava (HOS)** (Dept. of Computer Science & Engineering). This project report has been submitted as the proof of original work for the partial fulfillment of the requirement for the award of the degree of **Bachelor of Technology** (B. Tech) in the **Department of Computer Science & Engineering.** It has not been submitted anywhere else, under any other program to the best of our knowledge and belief.

|  |  |  |
| --- | --- | --- |
| **Team Members:** | | **Signatures:** |
| **(16CS005)** | **ANUJ PAREEK** | |
| **(16CS019)** | **JINY JAIN** | |
| **(16CS023)** | **MANISH RAWAT** | |
| **(16CS047)** | **VISHWAROOP SHAH** | |

**ACKNOWLEDGMENT**

A project of such a vast coverage cannot be realized without help from numerous sources and people in the organization. We take this opportunity to express our gratitude to all those who have been helping us in making this project successful.

We are highly indebted to our guide **Prof. Anubhav Saxena.** he has been a guide, motivator & source of inspiration for us to carry out the necessary proceedings for the project to be completed successfully. We also thank our project coordinator **Mr. Neeraj Prakash Shrivastava** for his co-operation, encouragement, valuable suggestions and critical remarks that galvanized our efforts in the right direction.

We would also like to convey our sincere thanks to **Mr. N.P. Shrivastava,** HOD, Department of Computer Science & Engineering, for facilitating, motivating and **s**upporting us during each phase of development of the project. Also, we pay our sincere gratitude to all the **Faculty Members** of Anand International College of Engineering Jaipur and all our Colleagues for their co-operation and support.

Finally, we would like to thank all those who have directly or indirectly helped and cooperated in accomplishing this project.

**Team Members:**

|  |
| --- |
| **ANUJ PAREEK (16CS005)** |
| **JINY JAIN (16CS019)** |
| **MANISH RAWAT (16CS023)** |
| **VISHWAROOP SHAH (16CS047)** |

**INDEX**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | **UNIT** | **DESCRIPTION** | | | |  | **PAGE No.** | | I. | TITLE PAGE | | | | | - | | II. | CERTIFICATE | | | | | 1 | | III. | ABSTRACT | | | | | 2 | | IV. | DECLARATION | | | | | 3 | | V. | ACKNOWLEDGEMENT | | | | | 4 | | 1. | PROJECT CHARTER | | | | | 23 | |  | 1.1 | Problem Statement & Objective | | | | 23 | |  | 1.2 | Literature Survey **/**Market Survey**/**Investigation & Analysis | | | | 23 | |  | 1.3 | Introduction to Project | | | | 25 | |  | 1.4 | Proposed Logic **/** Algorithm **/** Business Plan **/** Solution **/** Device | | | | 26 | |  | 1.5 | Scope of the Project | | | | 27 | | 2. | SYSTEM REQUIREMENT SPECIFICATION | | | | | 28 | |  | 2.1 | Overall Description | | | | 28 | |  |  | 2.1.1 | Product Perspective | | | 28 | |  |  |  | 2.1.1.1 | System Interfaces | | 28 | |  |  |  | 2.1.1.2 | User Interfaces | | 28 | |  |  |  | 2.1.1.3 | Hardware Interfaces | | 29 | |  |  |  | 2.1.1.4 | Software Interfaces | | 30 | |  |  |  | 2.1.1.5 | Communiations Interfaces | | 31 | |  |  |  | 2.1.1.6 | Memory Constraints | | 32 | |  |  |  | 2.1.1.7 | Operations | | 32 | |  |  |  | 2.1.1.8 | Site Adaptations Requirements | | 32 | |  |  | 2.1.2 | Project Functions | | | 32 | |  |  | 2.1.3 | User Characteristics | | | 33 | |  |  | 2.1.4 | Constraints | | | 33 | |  |  | 2.1.5 | Assumptions & Dependencies | | | 33 | |  | 2.2 | Specific Requirements | | | | 33 | |  |  | 2.2.1 | User Interface Requirements | | | 33 | |  |  | 2.2.2 | System Product Features | | | 34 | |  |  |  | 2.2.2.1 | Security | | 34 | |  |  |  | 2.2.2.2 | Maintainability | | 34 | |  |  |  | 2.2.2.3 | Portability | | 34 | | 3 | SYSTEM DESIGN SPECIFICATION | | | | | 35 | |  | 3.1 | System Architecture | | | | 35 | |  | 3.2 | Module Decomposition Description | | | | 35 | |  | 3.3 | High Level Design Diagrams | | | | 35 | |  |  | 3.3.1 | Usecase Diagram | | | 35 | |  |  | 3.3.2 | Activity Diagram | | | 35 | |  |  | 3.3.3 | Sequence Diagram | | | 36 | |  |  | 3.3.4 | Data-Flow Diagram | | | 37 | |  |  | 3.3.5 | Class Diagram | | | 38 | |  |  | 3.3.6 | Entity Relationship Diagram | | | 38 | |  |  | 3.3.7 | PERT Chart | | | 39 | | 4. | METHODOLOGY & TEAM | | | | | 40 | |  | 4.1 | Introduction to Waterfall Framework | | | | 41 | |  | 4.2 | Team Members, Roles & Responsibilities | | | | 42 | | 5. | SYSTEM TESTING | | | | | 51 | |  | 5.1 | Functionality Testing | | | | 51 | |  | 5.2 | Performance Testing | | | | 52 | |  | 5.3 | Usability Testing | | | | 53 | |  | 5.4 | Server-Side Interfacing | | | | 53 | |  | 5.5 | Client-Side Compatibility | | | | 53 | | 6. | TEST EXECUTION SUMMARY | | | | | 54 | | 7. | PROJECT SCREENSHOTS | | | | | 54 | | 8. | PROJECT SUMMARY AND CONCLUSIONS | | | | | 55 | | 9. | FUTURE SCOPE | | | | | 55 | | 10. | REFERENCES | | | | | 56 | | | | | |
|  | **INDEX OF FIGURES** | |  |
| **S. No** | **DESCRIPTION** |  | **PAGE No.** |
|  | TITLE PAGE | | 01 – 18 |
|  | CERTIFICATE | | 1 |
|  | ABSTRACT | | 2 |
|  | DECLARATION | | 3 |
|  | ACKNOWLEDGEMENT | | 4 |
|  | PROJECT CHARTER | | 5 |

|  |  |  |
| --- | --- | --- |
| **ABBREVIATIONS USED** | | |
|  | STLSD | Switching Theory & Logic System Design |
|  | MIS | Management Information System |
|  | DBMS | Database Management System |
|  | OS | Operaiting System |
|  | ANN | Artificial Neural Network |
|  | DMW | Data Mining & Warehousing |

**UNIT 1 - TECHNOLOGY OVERVIEW**

**1. Contributing Technologies**

**1.1 Machine Learning:**

Machine learning, and statistics are part of data science. The word learning in machine learning means that the algorithms depend on some data, used as a training set, to fine-tune some model or algorithm parameters. This encompasses many techniques such as regression, naive Bayes, classification or supervised clustering.

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.

Machine learning focuses on the development of computer programs that can access data and use it learn for themselves and Machine learning used for the prediction on basis of data.

**1.2 Type of Machine Learning:**

1. Supervised Machine Learning
2. Unsupervised Machine Learning
3. Reinforcement learning
4. Semi-supervised learning

Supervised and unsupervised learning describe two ways in which machines algorithms can be set loose on a data set and expected to learn something useful from it.

**1.2.1** **SUPERVISED MACHINE LEARNING:**

When an algorithm learns from example data and associated target responses that can consist of numeric values or string labels, such as classes or tags, in order to later predict the correct response when posed with new examples comes under the category of Supervised learning. This approach is indeed like human learning under the supervision of a teacher. The teacher provides good examples for the student to memorize, and the student then derives general rules from these specific examples.

1. Most of the practical machine learning uses supervised learning.
2. Supervised learning is where we have input variables (x) and an output variable (Y) and we use an algorithm to learn the mapping function from the input to the output.

Y = f(X)

1. The goal is to approximate the mapping function so well that when we have new input data (x) that we can predict the output variables (Y) for that data.
2. This process is called Supervised Machine learning.
3. Supervised learning problems can be further grouped into two problems.

**a) Classification**: A classification problem is when the output variable is a category, such as “red” or “blue” or “disease” and “no disease”.

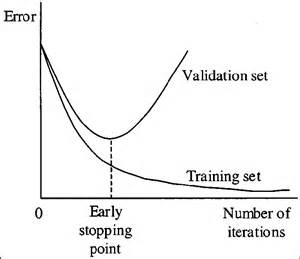
**b) Regression**: A regression problem is when the output variable is a real value, such as “dollars” or “weight”.

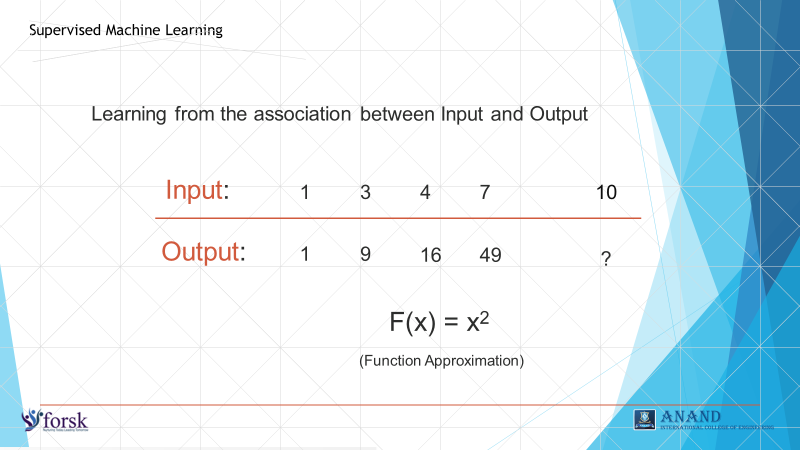
**Terms frequently used in this post**:

**Training Set:** a set of examples used for learning, where the target value is known.

**Validation Set:** asset of examples used to tune the architecture of a classifier and estimate the error.

**Test Set:** used only to assess the performances of a classifier. it is never used during the training process so that the error on the test set provides an unbiased estimate of the generalization error.





**Figure. 1.1** **Supervised Machine Learning**

**Regression:**

 Regression is useful for predicting outputs that are continuous. That means the answer to your question is represented by a quantity that can be flexibly determined based on the inputs of the model rather than being confined to a set of possible labels. Regression problems with time-ordered inputs are called time-series forecasting problems, like ARIMA forecasting, which allows data scientists to explain seasonal patterns in sales, evaluate the impact of new marketing campaigns, and more.

**Regression Algorithms**

* Simple linear Regression
* Multiple linear Regression
* Polynomial Regression
* Decision Tree Regression
* Random Forest Regression

**1.2.2 Unsupervised Machine learning**

If we are training our machine-learning task only with a set of inputs, it is called unsupervised learning, which will be able to find the structure or relationships between different inputs. Most important unsupervised learning is clustering, which will create different cluster of inputs and will be able to put any new input in appropriate cluster.

When an algorithm learns from plain examples without any associated response, leaving to the algorithm to determine the data patterns on its own. This type of algorithm tends to restructure the data into something else, such as new features that may represent a class or a new series of un-correlated values. They are quite useful in providing humans with insights into the meaning of data and new useful inputs to supervised machine learning algorithms  
  
As a kind of learning, it resembles the methods humans use to figure out that certain objects or events are from the same class, such as by observing the degree of similarity between objects. Some recommendation systems that you find on the web in the form of marketing automation are based on this type of learning.

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory[data mining,](https://en.wikipedia.org/wiki/Data_mining) and a common technique for[statistical](https://en.wikipedia.org/wiki/Statistics)[data analysis,](https://en.wikipedia.org/wiki/Data_analysis) used in many fields, including[machine learning,](https://en.wikipedia.org/wiki/Machine_learning)[pattern recognition,](https://en.wikipedia.org/wiki/Pattern_recognition)[image analysis,](https://en.wikipedia.org/wiki/Image_analysis)[information retrieval,](https://en.wikipedia.org/wiki/Information_retrieval)[bioinformatics,](https://en.wikipedia.org/wiki/Bioinformatics)[data compression,](https://en.wikipedia.org/wiki/Data_compression) and[computer graphics.](https://en.wikipedia.org/wiki/Computer_graphics)

**1.2.3.** **Reinforcement learning:**

When you present the algorithm with examples that lack labels, as in unsupervised learning. However, you can accompany an example with positive or negative feedback according to the solution the algorithm proposes comes under the category of Reinforcement learning, which is connected to applications for which the algorithm must make decisions (so the product is prescriptive, not just descriptive, as in unsupervised learning), and the decisions bear consequences. In the human world, it is just like learning by trial and error.

Errors help you learn because they have a penalty added (cost, loss of time, regret, pain, and so on), teaching you that a certain course of action is less likely to succeed than others. An interesting example of reinforcement learning occurs when computers learn to play video games by themselves.

In this case, an application presents the algorithm with examples of specific situations, such as having the gamer stuck in a maze while avoiding an enemy. The application lets the algorithm know the outcome of actions it takes, and learning occurs while trying to avoid what it discovers to be dan-gerous and to pursue survival. You can have a look at how the company Google DeepMind has created a reinforcement learning program that plays old Atari’s videogames. When watching the video, notice how the program is initially clumsy and unskilled but steadily improves with training until it becomes a champion.

**1.2.4** **Semi-supervised learning:**

where an incomplete training signal is given: a training set with some (often many) of the target outputs missing. There is a special case of this principle known as Transduction where the entire set of problem instances is known at learning time, except that part of the targets are missing.

**1.3 Some Popular Machine Learning Algorithms**

**1.3.1 Decision Trees**

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance-event outcomes, resource costs, and utility.

It is a type of supervised learning algorithm that is mostly used for classification problems. Surprisingly, it works for both categorical and continuous dependent variables. In this algorithm, we split the population into two or more homogeneous sets. This is done based on most significant attributes/ independent variables to make as distinct groups as possible.

In the image above, you can see that population is classified into four different groups based on multiple attributes to identify ‘if they will play or not’. To split the population into different heterogeneous groups, it uses various techniques like Gini, Information Gain, Chi-square, entropy.

From a business decision point of view, a decision tree is the minimum number of yes/no questions that one must ask, to assess the probability of making a correct decision, most of the time. As a method, it allows you to approach the problem in a structured and systematic way to arrive at a logical conclusion.

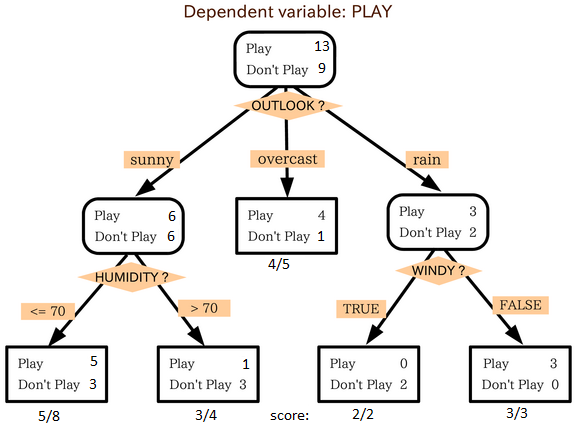
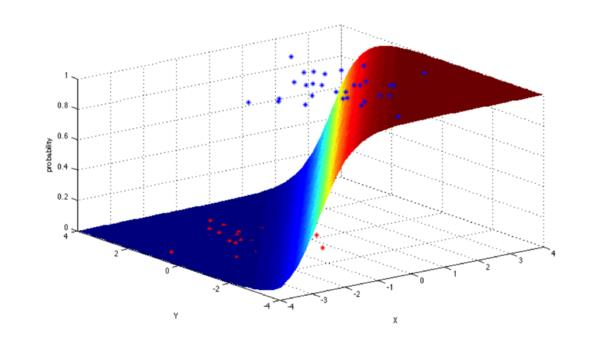
[](https://www.analyticsvidhya.com/wp-content/uploads/2015/08/IkBzK.png)

Figure 1.2: Decision Tree

**1.3.2 Logistic Regression**

Logistic regression is a powerful statistical way of modelling a binomial outcome with one or more explanatory variables. It measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution.

Don’t get confused by its name! It is a classification not a regression algorithm. It is used to estimate discrete values (Binary values like 0/1, yes/no, true/false) based on given set of independent variables(s). In simple words, it predicts the probability of occurrence of an event by fitting data to a logic function. Hence, it is also known as logit regression. Since, it predicts the probability, its output values lie between 0 and 1.



**Figure 1.3: Logistic Regression**

**1.3.3. Linear Regression**

It is used to estimate real values (cost of houses, number of calls, total sales etc.) based on continuous variable(s). Here, we establish relationship between independent and dependent variables by fitting a best line. This best fit line is known as regression line and represented by a Linear equation Y = a \* X + b.

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/08/Linear_Regression.png)

**Figure 1.4: Linear Regression**

In this equation:

* Y – Dependent Variable
* a – Slope
* X – Independent variable
* b – Intercept

These coefficients a and b are derived based on minimizing the sum of squared difference of distance between data points and regression line.

**1.3.4 KNN (K- Nearest Neighbors)**

It can be used for both classification and regression problems. However, it is more widely used in classification problems in the industry.

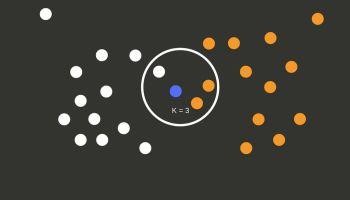
It is a lazy algorithm. What this means is that it does not use the training data points to do any generalization. K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases by a majority vote of its k neighbors. The case being assigned to the class is most common amongst its K nearest neighbors measured by a distance function. These distance functions can be Euclidean, Manhattan, Minkowski and Hamming distance. First three functions are used for continuous function and fourth one (Hamming) for categorical variables. If K = 1, then the case is simply assigned to the class of its nearest neighbors. At times, choosing K turns out to be a challenge while performing KNN modelling.

**Things to consider before selecting kNN:**

1. KNN is computationally expensive

2. Variables should be normalized else higher range variables can bias it

3. Works on pre-processing stage more before going for kNN like outlier, noise removal



**Figure 1.5: KNN**

**1.3.5 Naïve Bayes**

Naive Bayes is a collection of classification algorithms based on Bayes Theorem. It is not a single algorithm but a family of algorithms that all share a common principle, that every feature being classified is independent of the value of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3″ in diameter. A Naive Bayes classifier considers each of these “features” (red, round, 3” in diameter) to contribute independently to the probability that the fruit is an apple, regardless

of any correlations between features. Features, however, aren’t always independent which is often seen as a shortcoming of the Naive Bayes algorithm and therefore it’s labelled “naive”.

Although it’s a relatively simple idea, Naive Bayes can often outperform other more sophisticated algorithms and is extremely useful in common applications like spam detection and document classification.

In a nutshell, the algorithm allows us to predict a class, given a set of features using probability. So, in another fruit example, we could predict whether a fruit is an apple, orange or banana (class) based on its color, shape etc.(features).

**1.3.6 XGBoost (Extreme Gradient Boosting)**

XGBoost is an implementation of Gradient Boosting Machines (GBM) and is used for supervised learning. XGBoost (eXtreme Gradient Boosting) can be used for supervised learning tasks such as Regression, Classification, and Ranking. It is built on the principles of gradient boosting framework and designed to “push the extreme of the computation limits of machines to provide a scalable, portable and accurate library.”

**Gradient boosting:**

It is Machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

XGBoost is one of the implementations of Gradient Boosting concept, but what makes XGBoost unique is that it uses “a more regularized model formalization to control over-fitting, which gives it better performance. Therefore, it helps to reduce overfitting.

**1.4. Machine Learning Libraries**

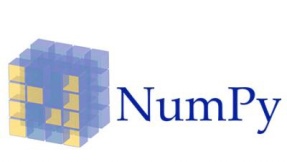
**1.4.1 About** **Libraries:**

Machine Learning, as the name suggests, is the science of programming a computer by which they can learn from different kind of data. A more general definition given by Arthur Samuel is – “Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed.” They are typically used to solve various types of life problems.

In the older days, people used to perform Machine Learning tasks by manually coding all the algorithms and mathematical and statistical formula. This made the process time consuming, tedious and inefficient. But in the modern days, it is become very much easy and efficient compared to the olden days by various python libraries, frameworks, and modules. 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
* Scikit-learn
* Theano
* TensorFlow
* Keras
* PyTorch
* Pandas
* Matplotlib

**1.4.1 NumPy**:  
NumPy is a very popular python library for large multi-dimensional array and matrix processing, with the help of a large collection of high-level mathematical functions. It is very useful for fundamental scientific computations in Machine Learning. It is particularly useful for linear algebra, Fourier transform, and random number capabilities. High-end libraries like TensorFlow uses NumPy internally for manipulation of Tensors.



**1.4.2 SciPy:**

SciPy is a very popular library among Machine Learning enthusiasts as it contains different modules for optimization, linear algebra, integration and statistics. There is a difference between the SciPy library and the SciPy stack. The SciPy is one of the core packages that make up the SciPy stack. SciPy is also very useful for image manipulation.



**1.4.3 Scikit-learn:**

Skikit-learn is one of the most popular ML libraries for classical ML algorithms. It is built on top of two basic Python libraries, viz., NumPy and SciPy. Scikit-learn supports most of the superised and unsupervised learning algorithms. Scikit-learn can also be used for data-mining and data-analysis, which makes it a great tool who is starting out with ML.



**1.4.4 Theano:**

We all know that Machine Learning is basically mathematics and statistics. Theano is a popular python library that is used to define, evaluate and optimize mathematical expressions involving multi-dimensional arrays in an efficient manner. It is achieved by optimizing the utilization of CPU and GPU. It is extensively used for unit-testing and self-verification to detect and diagnose different types of errors. Theano is a very powerful library that has been used in large-scale computationally intensive scientific projects for a long time but is simple and approachable enough to be used by individuals for their own projects.

****

**1.4.5 TensorFlow:**

TensorFlow is a very popular open-source library for high performance numerical computation developed by the Google Brain team in Google. As the name suggests, Tensorflow is a framework that involves defining and running computations involving tensors. It can train and run deep neural networks that can be used to develop several AI applications. TensorFlow is widely used in the field of deep learning research and application.

****

**1.4.6 Keras:**

Keras is a very popular Machine Learning library for Python. It is a high-level neural networks API capable of running on top of TensorFlow, CNTK, or Theano. It can run seamlessly on both CPU and GPU. Keras makes it really for ML beginners to build and design a Neural Network. One of the best thing about Keras is that it allows for easy and fast prototyping.



#### **1.4.7 PyTorch**

PyTorch is a popular open-source Machine Learning library for Python based on Torch, which is an open-source Machine Learning library which is implemented in C with a wrapper in Lua. It has an extensive choice of tools and libraries that supports on Computer Vision, Natural Language Processing (NLP) and many more ML programs. It allows developers to perform computations on Tensors with GPU acceleration and also helps in creating computational graph



#### **1.4.8 Pandas:**

Pandas is a popular Python library for data analysis. It is not directly related to Machine Learning. As we know that the dataset must be prepared before training. In this case, Pandas comes handy as it was developed specifically for data extraction and preparation. It provides high-level data structures and wide variety tools for data analysis. It provides many inbuilt methods for groping, combining and filtering data.



**1.4.9 Matplotlib:**

Matplotlib is a very popular Python library for data visualization. Like Pandas, it is not directly related to Machine Learning. It particularly comes in handy when a programmer wants to visualize the patterns in the data. It is a 2D plotting library used for creating 2D graphs and plots. A module named pyplot makes it easy for programmers for plotting as it provides features to control line styles, font properties, formatting axes, etc. It provides various kinds of graphs and plots for data visualization, viz., histogram, error charts, bar chats, etc.,

It is a 2D plotting library used for creating 2D graphs and plots. A module named pyplot makes it easy for programmers for plotting as it provides features to control line styles, font properties, formatting axes, etc. It provides various kinds of graphs and plots for data visualization, viz., histogram, error charts, bar chats, etc.



**UNIT – 2 PROJECT CHARTER**

**2.1 Problem Statement& Objective**

|  |
| --- |
| **Problem:** Terrorism detection from Social media using machine learning.  **Reasons:** There are several reasons for the problem   * social media played a major role in creating impact among the people before every terror activity. * People easily get misleaded from the social media. * Social media sites do not restrict people to post inappropriate content. * The communication or posting about such suspicious activities over social media creates a vast unrest among the people.   **Solution:**Keeping in mind that the social media influences a vast majority of people it is important to  Make sure that people must not misuses it. People trend to believe the things which are present on the social media and terrorist uses this to create a environment of uneasiness among the people. The solution which we have is that the unappropraite content must be removed from the social media. By making the system automatic the manual effort can be removed because the content which are posted on the social media is in vast quatity and it is infeasible to check it manually. We have target mainly the photograph and content which are there on the social media. Our system is going to use the machine learning model which we have trained so that it is able to detect automatically the inappropriate content and then appropriate actions can be made by using this method a lot of things can be avoided at early stage. With our system we would be able to remove those suspicious activities and strong action can be taken against the source of it. Youth without knowing the authenticity of the information tend to spread it across themselves which misleads them. |
|  |
| **Objective:**Is to build a mechanism using machine learning which is capable to identify the suspicious activities so that appropriate action must be taken against them. Performing Sentiment analysis on the content will help us filter the content. |

**2.2Investigation & Analysis** (Consultancy, Industry & Hardware Track)

|  |
| --- |
| The motivation for this research is taken from recent studies which have demonstrated that terrorist maily uses social media website likes twitter ,facebook to target there audience. |
|  |
| Extremist and terrorist groups use the internet for a myriad of purposes, including the dissemination of propaganda, the recruitment of new members and the development of operational planning. Online activity is a critical part of almost every national security investigation. By 1999 nearly all known terrorist groups The State of the Art 10 had established a presence on the internet. Nevertheless, the extent to which the internet affects radicalisation into violence is contested.  The picture is less clear in respect to social media specifically. Detailed empirical research into how extremist and terrorist groups have reacted to the rise of social media is limited, but markedly growing. The shift from text-heavy traditional websites to social networks built around interactive forums allowing the sharing of mixed media (often where leaders posted stories and steered discussions) came in the mid 2000s. Recent analysis suggests that since the late 2000s activity has increasingly shifted to social media platforms.  Social media accounts for an increasing proportion of time spent on-line. On an average day, Facebook users spend 9.7 billion minutes on the site, share 4 billion pieces of content a day and upload 250 million photos. Facebook is further integrated with 7 million websites and apps. |
|  |
| According to Aaron Zelin, ‘it is only a matter of time before terrorists use Twitter and Instagram as part of ongoing operations’. Zelin charts an increase in activists using Twitter as a tool of communication, motivated perhaps by the need to appeal to a younger demographic that prefers this medium. A MEMRI report has documented the use of Instagram by al-Qaeda leaders to share images and quotes, glorify imprisoned fighters, and disseminate images of dead ‘martyrs’. The international prominence (and highly cited case study, although recently discontinued) of al-Shabaab’s Twitter accounts has been used by the group to present a professional and united image, obtain support from the Somalia diaspora, offer dialogue with supporters and rebut critics in real time. |
| From right-wing to al-Qaeda inspired extremism, social media may ‘lower the bar’ for participation, making the involvement of low- The State of the Art 11 level, semi-radicalised or previously disengaged individuals a new feature of transnational extremist conversations and movements. Although extremist forums are still dominated by Arabic language content, the opposite is true of Twitter feeds. According to Michel Juneau-Katsuya, social media is playing a growing role in reaching out to vulnerable young people: ‘a means of privileged communication…which excludes their family and isolates them with others who sympathise with their cause and [who] think in a similar fashion.’ |
| Social media platforms are believed to have helped extend the reach of hate groups more broadly. According to Christopher Wolf, the online world ‘has become a technology embraced by racists, antiSemites, homophobes and bigots of all kinds to spread their message of hate’. Holocaust deniers, the Identity Church, KKK Members, neo-Nazis and racist skinhead groups are all believed to be particularly active. Anders Breivik, for example, drew much inspiration and impetus from his interactions online, including from the new ‘counter-Jihad movement’ – an international collection of Islamophobic bloggers, which, according to Hope not Hate, comprise over 200 organisations worldwide. |
| the International Network Against Cyberhate has argued that over recent years ‘the amount of cyber hate has grown to enormous proportions’, with ‘Islam, Jews, lesbians and gays, blacks, Roma, liberals’ and ‘leftwingers’ representing the main targets of online abuse. It is of note that of all the referrals made by the UK’s counter-terrorism internet Referral Unit (which seeks material that glorifies terrorism and asks for its removal from internet service providers), Facebook, Twitter, Blogger and/or Blogspot were most frequently identified as the hosts of the problematic, referred material. |
| More generally, social media use is affecting other types of law enforcement activity: criminal organisations and gangs exploit the internet and social media. Well-organised and longstanding groups have stable social media presences, usually used for advertising their organisation, or in some cases ‘cyber banging’ – levying threats against rival groups, or individuals. Indeed, in November 2012, the British Justice Secretary announced a crackdown on the use of social media by criminals to intimidate witnesses. Additionally, the amount of personal information posted on social media has been shown to influence the risks of individuals to burglary. |
| It is important to note that both the technological infrastructure of social media and the way that this infrastructure is used changes quickly. Research suggests that users have increasingly become aware of the privacy risks and reacted by placing more of their social media content onto higher privacy settings with more restricted possible readerships. |

**2.3Introduction to Project**

|  |
| --- |
| A Terrorism Detection from social media will help us to identity the inappropriate content from the social media which will help us to take appropriate actions at early stage.It has been observed that social media played a major role in creating impact among the people before every terror activity. The communication or posting about such suspicious activities over social media creates a vast unrest among the people. People trend to believe it and it creates a environment of fear among the people. The majority of people which uses the social media is the youth and they easily got diverted through it. The terrorist uses this as a opportunity and they uses social media to create a environment of fear and uneasiness among the youth . They have a proper method of how to brain wash the mind of the youth and they are well aware that a majority of people spends there a lot of time on the internet so they uses it an opportunity. |
|  |
| We are going to create a model using machine learning which will help us to identify inappropriate content on social media. We have trained the model in such a way that it is capable to identify the treating commets like Blast , bombs, attack etc all this will be treated as a suspicious and then appropriate actions could be taken against them . We have also work on images which will be able to detect those image which have or creates a feeling of fear like a person wearing a mask with a gun and there is a explosion in the background , such type of content can be marked as suspicious and the appropriate action can be taken against them. |

**2.4Solution**

|  |
| --- |
| The model which is based on machine learning will have to be trained first .So for that we are going to make two model one model will be used to work on textual data and another model will work on the images. For the model which will works on text is first trained using a large dataset of the suspicious content and normal content as well , so that our model is able to classify on the new data weather it is suspecios or not. For this purpose we will use use the NLP (Natural Language Processing) it will create a bag of words. It mainly determine the frequency of the words which occur in both the suspicious and normal data so that when we give them a new data it will measure the occurrence of the words and then it will compare it with those which we have trained so in this way it will be able to classify the content. |
| In case of the model which will work on images. We are going to trained it first on the dataset which will contain both the positive and negative images. Positon images are those which have been fined as the suspicious activites and negative images are those images which are not find suspicious.  We have use the Convolutional Neural Network (CNN) which will be trained on that dataset.  The problem with images is that it is consist of pixels so we have have to work on pixels and since there is a lot of pixels on the image so if we trained on all then it will it take a lot of time .  Instead we are first going to do the feature selection on the images so that all the important featues are first gets collected in a image then our machine will be trained on that featues and not on all the images .It will create the more efficient model .  Int this way we are going to train both the models one using the Natural Language Processing(NLP) and another using the Convolutional Neural Network(CNN).  Image result for Working process of CNN  **Fig2.1 Working of Convolutional Neural Network (CNN)** |

**2.5Scope of the Project**

The project entitled “Terrorism detection from social media” will going to help us by providing the facility of automatically detecting the suspecious activities and then taking the appropriate steps at the right time. Because the content posted in social media is enormous and it will infeasible to check the content manually. The benefit from this project will be that the people will not be get diverted because the content will be removed if it contains suspicious content.

The system is going to help us to identify the content which are found to be suspicious. In todays world when all the things are coming on the internet the more users are being getting influenced by it. Our application is going to play a vital role in the coming time. At the time in our application we must load the content in our application to determine weather it is suspicious or not. In our next step what we will do is that we mould our application in such a way that it is able to automatically acquire the data from the site when it is loaded and identify weather it is suspicious or not. If found suspicious then appropriate actions can be taken against it.

The success criteria for the system are based in the level up to which the features described in this document are implemented in the system.

**UNIT – 3 SYSTEM REQUIREMENT SPECIFICATION**

**3.1 Overall Description**

This section and its subsections contain the description of the project compoents such as interfaces, performance requirements, design constraints, assumptions and dependencies etc.

**3.1.1Product Perspective**

The application will be a Windows / Linux based, self contained and independent product.

**3.1.1.1System Interfaces**

List each system interface and identify the functionality of the system (hardware and software both) to accomplish the system requirement and interface description to match the system.

**3.1.1.2User Interfaces**

|  |  |
| --- | --- |
|  | The application will have a user friendly and menu based interface. Following screens will be provided:   1. A webpage where user can load the image or a text. 2. Following output will be gven:  * Image: If the image is given as input then it will tell weather it is suspicious or not * Text: If the input is text then it will do sentiment analysis and tell weather it is suspicious or not. |

**3.1.1.3Hardware Interfaces**

* Screen resolution of at least 800 x 600 pixels is required for proper and complete viewing of screens. Higher resolutions in wide-screen mode will be better for a better view.
* A network connection (internet) is required to make the service accessible to the users
* Other hardware interface specifications are as follows

|  |  |
| --- | --- |
| **HARDWARE INTERFACES - CLIENT SIDE (Minimum)** | |
| **HARDWARE** | **SPECIFICATIONS** |
| Snapdragon 750 /MediaTek Helio | NONE SUCH |
| 3 GB Ram | NONE SUCH |
| Qwerty keyboard or Touch screen | NONE SUCH |
| GPS / NavIC Chips | NONE SUCH |
| 5MP Camera | NONE SUCH |

**Table 3.1 – Minimum Client-Side Hardware Interfaces**

|  |  |  |
| --- | --- | --- |
| **HARDWARE INTERFACES - SERVER SIDE (Minimum)** | | |
| **HARDWARE** | **RAM** | **DISK SPACE** |
| Intel Core i3 / i5 / i7 2.60 GHz and higher  Or  AMD Ryzen/ Threadripper | 16 GB | 10 TB |

**Table 3.2 – Minimum Server-Side Hardware Interfaces**

|  |  |  |
| --- | --- | --- |
| **HARDWARE INTERFACES - CLIENT SIDE (Recommended)** | | |
| **HARDWARE** | **RAM** | **DISK SPACE** |
| Snapdragon 820 /MediaTek Helio | 4GB | 1 GB |
| Qwerty keyboard or Touch screen |
| 12 MP Camera |
| GPS / NavIC Chips |

**Table 3.3 – Recommended Client-Side Hardware Interfaces**

|  |  |  |
| --- | --- | --- |
| **HARDWARE INTERFACES - SERVER SIDE (Recommended)** | | |
| **HARDWARE** | **RAM** | **DISK SPACE** |
| Intel Xeon higher  AMD Ryzen/ Threadripper | 32GB | 40 GB |

**Table 3.4 – Recommended Server-Side Hardware Interfaces**

**3.1.1.4Software Interfaces**

* Any Microsoft Windows 7 and higher (Windows 7 / 8 / 8.1 / 10) or equivalent Linux based operating systemwith minimum kernel support 3.X.
* Crystal Reports 8 for generation and viewing of reports
* J2EE (JSP, Java Bean, Servlets, HTML, XML, AJAX)& jQuery for coding and developing of the application.

|  |  |  |
| --- | --- | --- |
| **SOFTWARE INTERFACES (Minimum)** | | |
| **Software Tool** | **Version** | **Purpose of Use** |
| Operating system | Windows 7 and higher  or Linux with Kernel 3.x and higher | Installation and operational platform |
| Web Browser | Internet Explorer 11 and other higher compatible | Access to the web application |
| Web Server | Google Cloud Plateform | Running the web application over internet / intranet |
| Database | Firebase | Running and linking the database over internet / intranet to the online web application |

**Table 3.5 – Minimum Software Interfaces**

|  |  |  |
| --- | --- | --- |
| **SOFTWARE INTERFACES (Recommended)** | | |
| **Software Tool** | **Version** | **Purpose of Use** |
| Operating system | Windows 10 &higher  or Linux with Kernel 4.x & higher | Installation and operational platform |
| Web Browser | Chrome 78 or Firefox 72 | Access to the web application |
| Web Server | Google Cloud Plateform | Running the web application over internet / intranet |
| Database | Firebase | Running and linking the DB on internet / intranet |

**Table 3.6 – Recommended Software Interfaces**

**3.1.1.5Communiation Interfaces**

* Client (customer) on Internet will be using HTTP/HTTPS protocol.
* Client (system user) on Internet will be using HTTP/HTTPS protocol.

**3.1.1.6Memory Constraints**

* At least 2 GB of RAM and 1 GB of space on hard disk will be required for running the application on client end.
* Similarily, a minimum of 16 GB of RAM and 200 GB of space on hard disk will be required for running the application on server end.

**3.1.1.7Operations**

* This product release will not cover any automated housekeeping aspects of the database.
* The DBA at the client side will be assumed responsible for manually deleting or archieving obsolete or non-required data from the database as per client’s requirements.
* This will include database backup and recovery options also.
* The Apache Tomacat webserver will be hosted and maintained on a remote server addressed by a URL based address.
* The URL address may be intranet or internet based as per client’s requirements.
* The ‘SYSTEM RESET’ function is provided that after confirmation from the administrator, will delete all the selective or complete data from the system.

**3.1.1.8Application Adaption Requirements**

The computing terminals conneted to network (internet) at the client end will be required to support the hardware and software interfaces specified in above sections.

**3.1.2Project Functions**

The system will allow access only to authorized uesrs with specific roles (Administrator, user, guest). Depending upon the user’s role, he / she will be able to access only specific modules of the system.

A summary of the major functions that the software will perform:

1. User (with role Data Entry Operator) will able to add / modify / delete information about different terror attacke held in the country.
2. User (with role Co-ordinator) will be able to generate printable reports.
3. Users (with role Administrator) will be able to ‘reset’ the system-leading to deletion of all existing information from the backend database.

Users (with role Administrator) will be able to create / modify / delete new / existing user accounts.

**3.1.3User Characteristics**

* *Educational Level:* User should be comfortable with Hindi or English.
* *Experience:* User should be able to capture and make the small clipps and transfer it to the emergency system portal and able to cheak the status of the emergency services.
* *Technical Expertise:* User should be comfortable using general purpose applications on a mobile.

**3.1.5Assumptions & Dependencies**

* The number of domains being selected by the user does not change.
* The subject types (*i.e. terror detection and its minimization)* do not change.

**3.2Specific Requirements**

This section presents the software requirements to a level of detail sufficiency to enable designers to design and testers to test the system.

**3.2.1User Interface Requirements**

Following screens will be provided by the system:

**Home Panel:**

This screen will be accessible only to users at client-side. It will allow the user to enter the text in the input field and to see the output generated by the system.

**Contact Us:**

For any help and suggestions regarding the applications users are free to contact the administrator using the details provided on this page.

**About us:**

This page is user accessible where the details of individual programmer is maintained.

**3.2.2System Product Features**

**3.2.2.1Security**

The application will be tested properly before being deployed on internet. Users will have to enter correct address to reach the website.

**3.2.2.2Maintainability**

The application will be designed in a manner to make it easy to incorporate new requirements in individual modules such as further extention of the application where Images, Videos and Audios will be analysed to predict the output.

**3.2.2.3Portabilty**

The application will be easily portable among any windows/linux or smartphones.

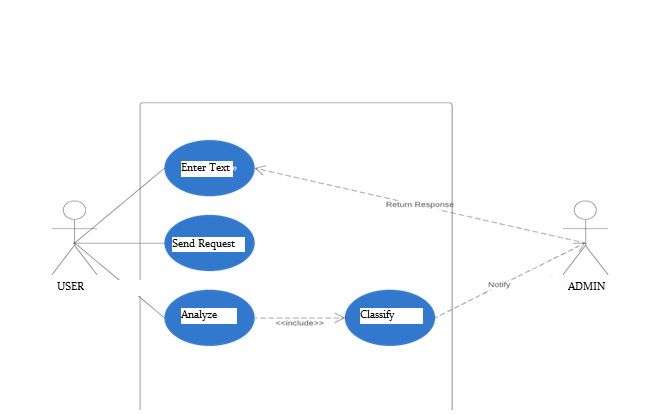
**UNIT– 4 SYSTEM DESIGN SPECIFICATION**

**4.1System Architecture**

System architecture presents the schematic view of the complete system along with its major components and their connectivities. The overall architecture of the proposed system will be as follows.

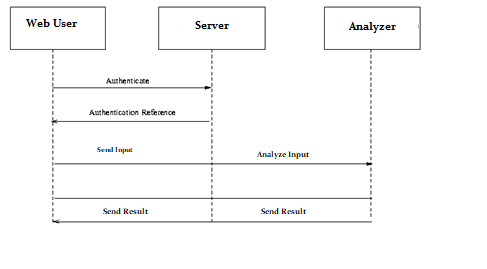
**4.2High Level Design Diagrams**

**4.2.1Usecase Diagrams**



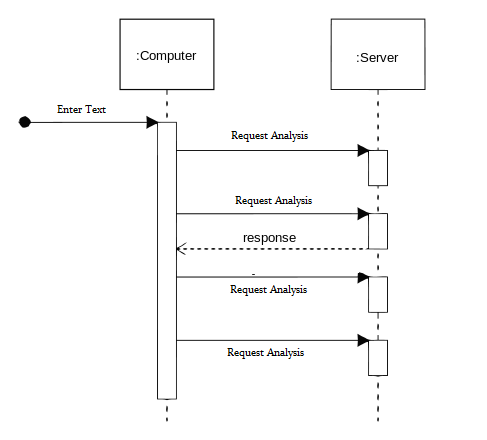
**Figure 4.1 – Usecase Diagram**

**4.2.2Activity Diagram**



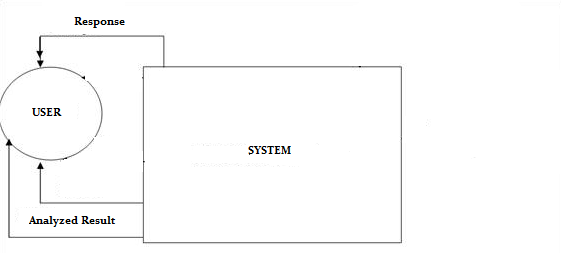
**Figure 4.2–ActivityDiagram**

**4.2.3Sequence Diagram**

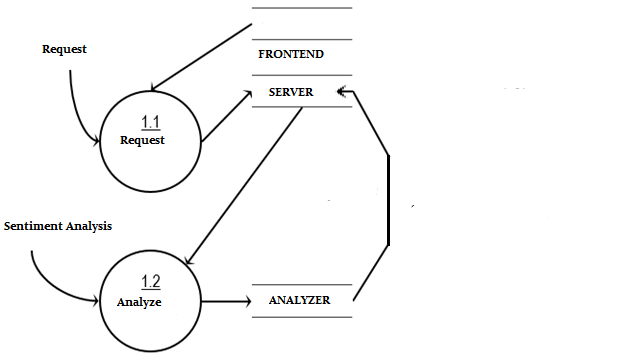


**Figure 4.3–SequenceDiagram**

**4.2.4Data-Flow Diagram**

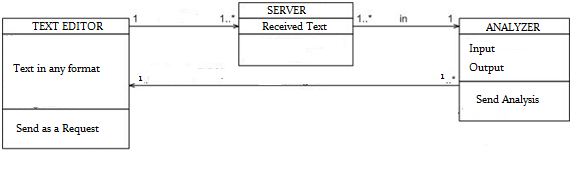


**Figure 4.4–Level0 Dataflow Diagram**



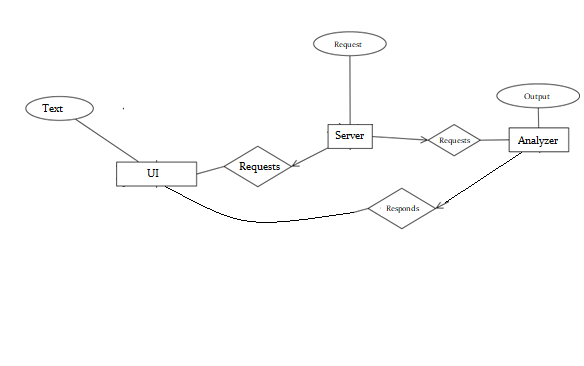
**Figure 4.5 – Level 1 Dataflow Diagram**

**4.2.5Class Diagram**



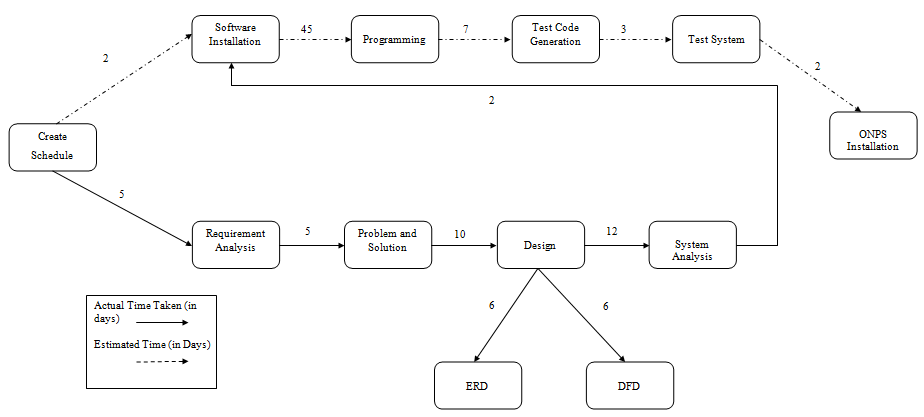
**Figure 4.6 – Class Diagram**

**4.2.6Entity Relationship Diagram**



**Figure .7 – Entity Relationship Diagram**

**4.2.7PERT Chart**



**Figure 4.8 – P.E.R.T. Chart**

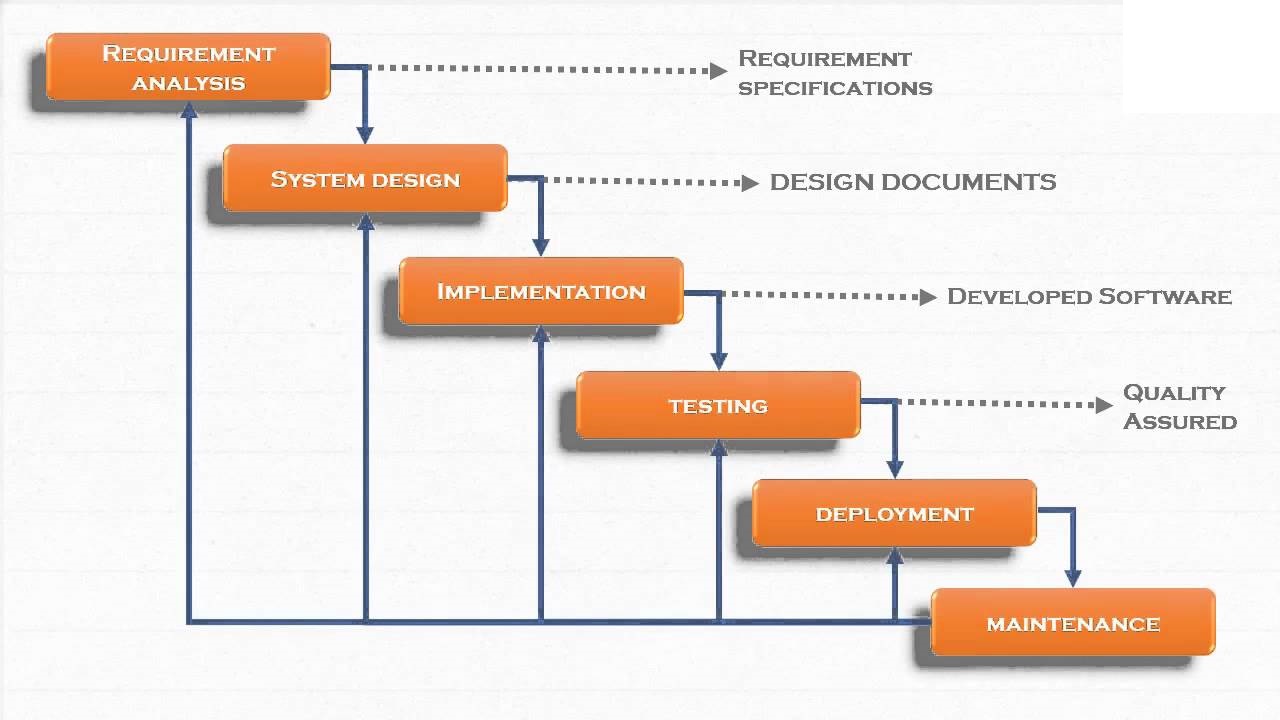
**UNIT– 5 METHODOLOGY & TEAM**

**5.1Introduction to Waterfall Model**

The Waterfall Model was first Process Model to be introduced. It is also referred to as a linear-sequential life cycle model. It is very simple to understand and use. In a waterfall model, each phase must be completed before the next phase can begin and there is no overlapping in the phases.

The waterfall Model illustrates the software development process in a linear sequential flow; hence it is also referred to as a linear-sequential life cycle model. This means that any phase in the development process begins only if the previous phase is complete. In waterfall model phases do not overlap. In "The Waterfall" approach, the whole process of software development is divided into separate phases. In Waterfall model, typically, the outcome of one phase acts as the input for the next phase sequentially.

Following is a diagrammatic representation of different phases of waterfall model.



**Figure 5.1 – Waterfall Model with Feedback**

The sequential phases in Waterfall model are:

* **Requirement Gathering and analysis:** All possible requirements of the system to be developed are captured in this phase and documented in a requirement specification doc.
* **System Design:** The requirement specifications from first phase are studied in this phase and system design is prepared. System Design helps in specifying hardware and system requirements and also helps in defining overall system architecture.
* **Implementation:** With inputs from system design, the system is first developed in small programs called units, which are integrated in the next phase. Each unit is developed and tested for its functionality which is referred to as Unit Testing.
* **Integration and Testing:** All the units developed in the implementation phase are integrated into a system after testing of each unit. Post integration the entire system is tested for any faults and failures.
* **Deployment of system:** Once the functional and non functional testing is done, the product is deployed in the customer environment or released into the market.
* **Maintenance:** There are some issues which come up in the client environment. To fix those issues patches are released. Also to enhance the product some better versions are released. Maintenance is done to deliver these changes in the customer environment.

All these phases are cascaded to each other in which progress is seen as flowing steadily downwards (like a waterfall) through the phases. The next phase is started only after the defined set of goals are achieved for previous phase and it is signed off, so the name "Waterfall Model". In this model phases do not overlap.

Every software developed is different and requires a suitable SDLC approach to be followed based on the internal and external factors. Some situations where the use of Waterfall model is most appropriate are:

* Requirements are very well documented, clear and fixed.
* Product definition is stable.
* Technology is understood and is not dynamic.
* There are no ambiguous requirements.
* Ample resources with required expertise are available to support the product.
* The project is short.

## **Waterfall Model Pros & Cons**

### Advantage

The advantage of waterfall development is that it allows for departmentalization and control. A schedule can be set with deadlines for each stage of development and a product can proceed through the development process model phases one by one.

Development moves from concept, through design, implementation, testing, installation, troubleshooting, and ends up at operation and maintenance. Each phase of development proceeds in strict order.

### Disadvantage

The disadvantage of waterfall development is that it does not allow for much reflection or revision. Once an application is in the testing stage, it is very difficult to go back and change something that was not well-documented or thought upon in the concept stage.

**5.2Team Members, Roles & Responsibilities**

|  |  |
| --- | --- |
| **Team Member** | **Project Role** |
| **Anuj** | **Overall Coordination, NLP** |
| **Jiny** | **Documentation, Connections** |
| **Vishwaroop** | **Data Extraction and Analysis** |
| **Manish** | **Frontend designing & Documentation** |

**Table 5.1 – Roles and responsibilities**

**UNIT– 6 SYSTEM CODE**

**6.1 Projects Code: -**

The system uses Natural Language Processing to identify for which type of sentimemt does the text hold. Once the user enters the text and clicks on ‘ANALYZE’, The system predicts The sentiment of the text and shows it to the user.

We built this solution in the following steps: -

1. Data Gathering
2. Model Training
3. User-Interface
4. Setting up server
   * 1. **Data Gathering: -**

Data gathering was one of the biggest challenges in our project development phase. The text from different websites , comments from social media was collected and stored in a proper format.

**6.1.2Model Training: -**

After successfully gathering the data, we built a Natural Language Processing model in Python Language. We trained our model on Anaconda platform, which is a platform to develop and run the applications.

This is the code for our model training: -

# Natural Language Processing

# Importing the libraries

import pandas as pd

# Importing the dataset

# Ignore double qoutes, use 3

dataset = pd.read\_csv('FinalData.csv')

dataset=dataset.iloc[:400,:]

# Cleaning the texts

# Noise removal

""" language stopwords

(commonly used words of a language – is, am, the, of, in etc),

URLs or links, social media entities (mentions, hashtags),

punctuations and industry specific words.

This step deals with removal of all types of noisy entities present

in the text.

"""

import re

import nltk

nltk.download('stopwords')

from nltk.corpus import stopwords

#Stemming: Stemming is a rudimentary rule-based process

# of stripping the suffixes (“ing”, “ly”, “es”, “s” etc) from a word.

"""

The most common lexicon normalization practices are :

Stemming: Stemming is a rudimentary rule-based process of stripping the

suffixes (“ing”, “ly”, “es”, “s” etc) from a word.

Lemmatization: Lemmatization, on the other hand, is an

organized

& step by step procedure of obtaining the root form of

the word,

it makes use of vocabulary (dictionary importance of words)

and

morphological analysis (word structure and grammar relations).

"""

from nltk.stem.porter import PorterStemmer

#from nltk.stem.wordnet import WordNetLemmatizer

corpus = []

#perform row wise noise removal and stemming

'''

#let's do it on just first row data

phrase = re.sub('[^a-zA-Z]', ' ', dataset['phrase'][0])

phrase = phrase.lower()

phrase = phrase.split()

phrase = [word for word in phrase

if not word

in set(stopwords.words('english'))]

#lem = WordNetLemmatizer() #Another way of finding root word

ps = PorterStemmer()

phrase = [ps.stem(word) for word in phrase]

#phrase = [lem.lemmatize(word) for word in phrase if not word in set(stopwords.words('english'))]

phrase = ' '.join(phrase)

corpus.append(phrase)

'''

#now do the same for every row in dataset. run to loop for all rows

for i in range(0, len(dataset.phrase)):

phrase = re.sub('[^a-zA-Z]', ' ', dataset['phrase'][i])

phrase = phrase.lower()

phrase = phrase.split()

phrase = [word for word in phrase if not word in set(stopwords.words('english'))]

#lem = WordNetLemmatizer() #Another way of finding root word

ps = PorterStemmer()

phrase = [ps.stem(word) for word in phrase]

#phrase = [lem.lemmatize(word) for word in phrase if not word in set(stopwords.words('english'))]

phrase = ' '.join(phrase)

corpus.append(phrase)

# Creating the Bag of Words model

# Also known as the vector space model

# Text to Features (Feature Engineering on text data)

from sklearn.feature\_extraction.text import CountVectorizer

cv = CountVectorizer(max\_features = 5000)

features = cv.fit\_transform(corpus).toarray()

labels = dataset.iloc[:, 1].values

'''

# Splitting the dataset into the Training set and Test set

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

features\_train, features\_test, labels\_train, labels\_test = train\_test\_split(features, labels, test\_size = 0.20, random\_state = 0)

'''

"""

# Fitting Naive Bayes to the Training set

from sklearn.naive\_bayes import GaussianNB

classifier = GaussianNB()

classifier.fit(features\_train, labels\_train)

# Predicting the Test set results

labels\_pred = classifier.predict(features\_test)

# Making the Confusion Matrix

from sklearn.metrics import confusion\_matrix

cm\_nb = confusion\_matrix(labels\_test, labels\_pred)

"""

#applying knn on this text dataset

# Fitting Knn to the Training set

from sklearn.neighbors import KNeighborsClassifier

classifier = KNeighborsClassifier()

classifier.fit(features, labels)

'''

# Predicting the Test set results

#labels\_pred = classifier.predict(features\_test)

# Making the Confusion Matrix

from sklearn.metrics import confusion\_matrix

cm\_knn = confusion\_matrix(labels\_test, labels\_pred)

'''

input\_string=input("Enter the statement")

phrase = re.sub('[^a-zA-Z]', ' ', input\_string)

phrase = phrase.lower()

phrase = phrase.split()

phrase = [word for word in phrase

if not word

in set(stopwords.words('english'))]

#lem = WordNetLemmatizer() #Another way of finding root word

ps = PorterStemmer()

phrase = [ps.stem(word) for word in phrase]

#phrase = [lem.lemmatize(word) for word in phrase if not word in set(stopwords.words('english'))]

phrase = ' '.join(phrase)

corpus=[]

corpus.append(phrase)

inputPred = cv.transform(corpus).toarray()

* + 1. **User-Interface: -**

To make our application user friendly we made a simple and responsive user interface for which the code is shown below: -

For Index page: -

<!DOCTYPE html>

<html>

<head>

<title>TERRORISM DETECTION</title>

<link rel="stylesheet" type="text/css" href="f2.css">

<link href="https://fonts.googleapis.com/css?family=Raleway&display=swap" rel="stylesheet">

</head>

<body>

<h1 align="center">Terrorism Detection</h1>

<form action="/http://www.google.com" method="get">

<input type="text" name="textbox">

<button type="analyze" value="analyze">analyze</button>

<button type="reset" value "Reset">Reset</button>

<ul>

<li>This page analyzes text</li>

</ul>

<ol>

<li>Enter text in the text box</li>

<li>click on analyze</li>

<li>Click on reset to delete everything</li>

</ol>

</form>

</body>

</html>

**CSS for the webpage: -**

input[type="text"]

{

margin-top:80px;

border: 2px solid blue;

margin-left: 450px;

margin-right: 50px;

width: 400px;

height: 100px;

}

body

{

background-repeat: no-repeat;

background-size: cover;

}

h1{

padding-top: 25px;

padding-bottom: 25px;

background: rgb(0,0,0);

font-family:Rockwell;

color: rgba(255,255,255,1);

font-size: 30px;

border-bottom: 3px solid red;

border-color:black;

}

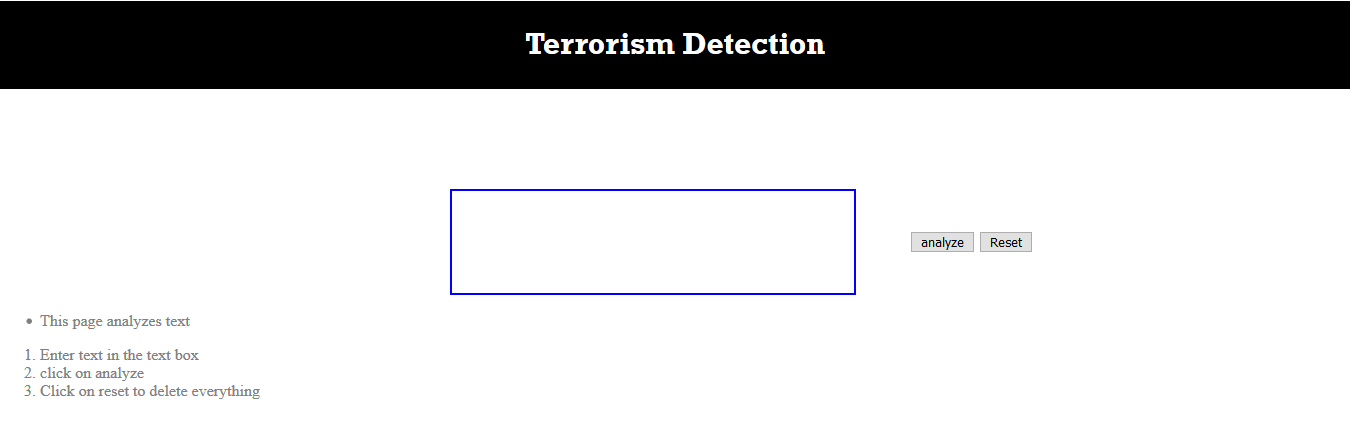
li{

color: grey;

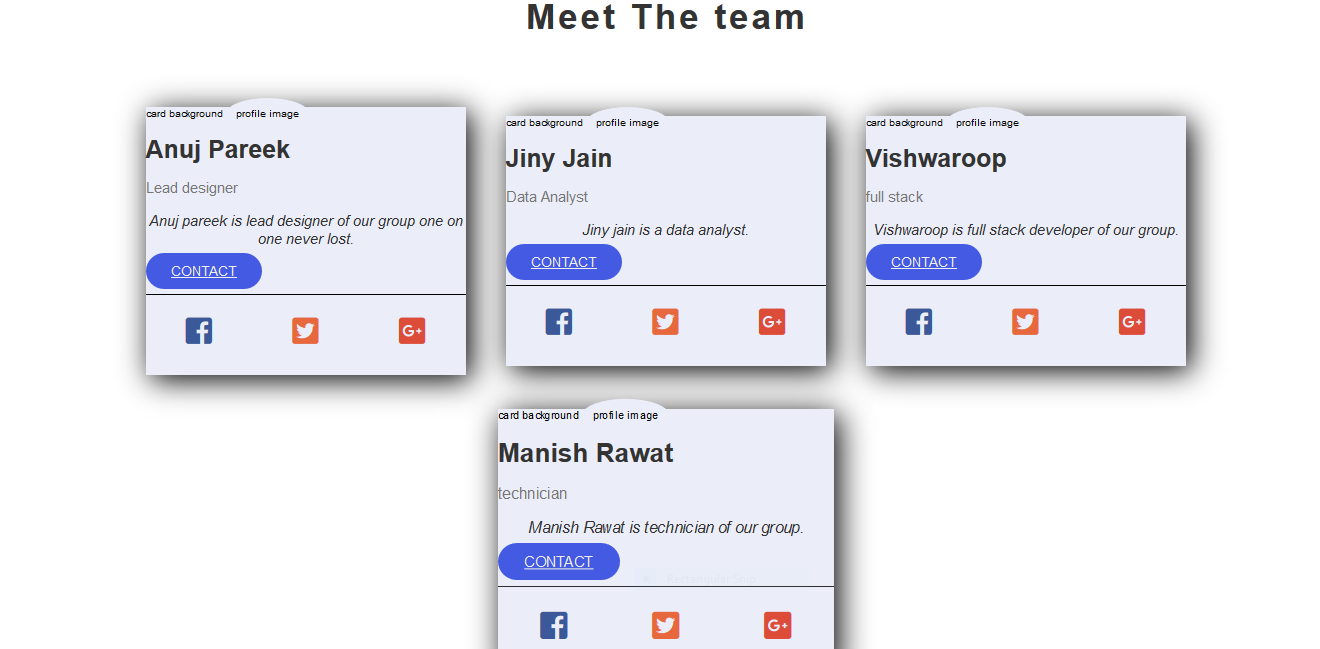
}

Our user-interface was built using HTML, CSS and Bootstrap.

Following are the screenshots of our Web-Application: -



**Fig:6.5 Homepage**



**Fig:6.6 Contact Us**

**UNIT– 7SYSTEM TESTING**

The designed system has been testing through following test parameters.

**7.1FunctinalityTesting**

FUNCTIONAL TESTING is a type of software testing whereby the system is tested against the functional requirements/specifications.

Functions (or features) are tested by feeding them input and examining the output. Functional testing ensures that the requirements are properly satisfied by the application. This type of testing is not concerned with how processing occurs, but rather, with the results of processing. It simulates actual system usage but does not make any system structure assumptions.

During functional testing, Black Box Testing technique is used in which the internal logic of the system being tested is not known to the tester.

Functional testing is normally performed during the levels of System Testing and Acceptance Testing.

Typically, functional testing involves the following steps:

* Identify functions that the software is expected to perform.
* Create input data based on the function’s specifications.
* Determine the output based on the function’s specifications.
* Execute the test case.
* Compare the actual and expected outputs.

Functional testing is more effective when the test conditions are created directly from user/business requirements. When test conditions are created from the system documentation (system requirements/ design documents), the defects in that documentation will not be detected through testing and this may be the cause of end-user’s wrath when they finally use the software.

**7.2Performance Testing**

Performance Testing is defined as a type of software testing to ensure software applications will perform well under their expected workload.

Features and Functionality supported by a software system is not the only concern. A software application's performance like its response time, reliability, resource usage and scalability do matter. The goal of Performance Testing is not to find bugs but to eliminate performance bottlenecks.

The focus of Performance Testing is checking a software program's

* Speed - Determines whether the application responds quickly
* Scalability - Determines maximum user load the software application can handle.
* Stability - Determines if the application is stable under varying loads

Performance Testing is popularly called “Perf Testing” and is a subset of performance engineering.

Performance Testing is done to provide stakeholders with information about their application regarding speed, stability, and scalability. More importantly, Performance Testing uncovers what needs to be improved before the product goes to market. Without Performance Testing, software is likely to suffer from issues such as: running slow while several users use it simultaneously, inconsistencies across different operating systems and poor usability.

Performance testing will determine whether their software meets speed, scalability and stability requirements under expected workloads. Applications sent to market with poor performance metrics due to nonexistent or poor performance testing are likely to gain a bad reputation and fail to meet expected sales goals.

Also, mission-critical applications like space launch programs or life-saving medical equipment should be performance tested to ensure that they run for a long period without deviations.

**7.3Usability Testing**

Usability Testing is defined as a type of software testing where, a small set of target end-users, of a software system, "use" it to expose usability defects. This testing mainly focuses on the user's ease to use the application, flexibility in handling controls and the ability of the system to meet its objectives. It is also called User Experience (UX) Testing.

This testing is recommended during the initial design phase of SDLC, which gives more visibility on the expectations of the users.

Aesthetics and design are important. How well a product looks usually determines how well it works.

There are many software applications/websites, which miserably fail, once launched, due to following reasons -

* Where do I click next?
* Which page needs to be navigated?
* Which Icon or Jargon represents what?
* Error messages are not consistent or effectively displayed
* Session time not enough.

Software Engineering, Usability Testing identifies usability errors in the system early in the development cycle and can save a product from failure.

**7.4 Server-Side Interfacing**

In this we tested the server-side interface. This was done by verifyingthatcommunication is done properly. Also, the compatibility of server with software, hardware, network and database was tested.

**7.5 Client-Side Compatibility**

The client-side compatibility is also tested using various browsers like Google Chrome, Mozilla Firefox and Internet Explorer.

**UNIT– 8 PROJECT SUMMARY AND CONCLUSIONS**

The object of this project is to harness the power of Internet for our practical and potential one. This report explains to extensively cover this concept and plant a seed of inquisitiveness in the mind of users.

This application people in carpooling which reduce the burden of travelling expensis for them as well as project will reduce the traffic which we see on roads and also reduces the enviormental pollution and reduces the usage of natural resources so they can be long lasting.

Our project is currently on track with the exception of a couple of problems that were encountered. The first is that we have not yet figured out how to. Implementation of this function will require further research.

The second problem is determining the reason why we can only. To troubleshoot this problem we must look over the code more thoroughly. We are still confident that we will be able to find solutions to these problems, so we expect not to have to compromise on our initial project goal.

**UNIT– 9FUTURE SCOPE**

The possible future scope of this application will be on mobile platform with following enhancements:

1. Designing of mobile application for this project on Android / iPhone platforms.

2. Working on pictures i.e analysis on images.

3. Analysis on videos.

4. Embedding this on a social media application.

**UNIT– 10REFERENCES**

**BOOKS AND PAPERS:**

## 1. [NLP at work, by Sue Knight](http://amzn.to/2m7BTlm)

## 2. Words that change minds, by Shelle Rose Charvet

3. Quick guide to NLP, by Brian Weiss

**WEBSITES:**

1. <http://www.stackoverflow.com>

2. <http://developers.google.com>

3. https://medium.com/@ritidass29/the-essential-guide-to-how-nlp-works-4d3bb23faf76