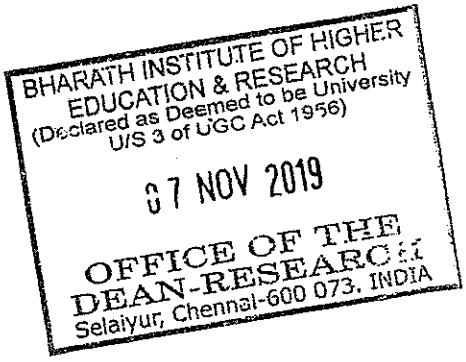
Bharath
INSTITUTE OF HIGHER EDUCATION AND RESEARCH
(Declared as Deemed-to-be University under section 3 of UGC Act, 1956)
(Vide Notification No. F.9-5/2000 - U.3, Ministry of Human Resource Development, Govt. of India, dated 4th July 2002)
Accredited by NAAC
Phone : 044-22290742 / 22290125 . Telefax : 044-22293886
Website : www.bharathuniv.ac.in
173, Agaram Road, Selaiyur, Tambaram,
Chennai - 600 073. Tamil Nadu.

Ref No.SMS-2018-O-29 Date: 07/11/2019



TO

Mr. Dr. S.Thiru Nirai Senthil,

Associate Professor/ CSE,

BIHER.

Thro: Concern Head of the Department

Greetings!!!

We are happy to announce that the Research Advisory Committee has approved your proposal for Seed Money Scheme-2018 which was presented by you. You are requested to complete the proposal and send the progress report to the Dean Research in the prescribed time period.

Title of the Project: Prediction of various diseased using sentimental analyses.

Seed Money Amount: Rs.1, 00,000/- (Rupees One Lakh Only)

Approved on: 04/11/2019

Payment details:

Voucher No.29 **Dated: 08/11/2019**

With Regards

Dean-Research

Bharath University

SELAIYUR, CHENNAI - 600 073, TAMIL NADU, INDIA.

CASH / PAYMENT VOUCHER

Date 06/11/18

V.No. : 029

Debit _____ Amount

Rs. 1,00,000/-

PAID TO Mr. S. Thirunirai Senthil

RUPEES one lakh only.

TOWARDS Seed Money Scheme - 2018.

BHARATH UNIVERSITY CHENNAI 600 073. INDIA

Authorised by **Finance Manager** **Cashier/Accountant** **Payee's Signature**

PROPOSAL SUBMISSION

1. Details of Principal Investigator

Name	: Dr. S. THIRU NIRAI SENTHIL
Designation	: Associate Professor
Highest Qualifications	: Ph.D
Department	: Computer Applications
E-mail	: senthil081074@gmail.com
Contact no	: 9443645536
Date of Joining	: 12.07.2017

2. Details of Co - Principal Investigator

Name	: Mr.S.Kannan
Designation	: Assistant Professor
Highest Qualifications	: M.C.A., M.Phil
Department	: Computer Science
E-mail	: ks14091979@gmail.com
Contact no	: 9940301564
Date of Joining	: 05.09.2003

Technical details

1. Introduction

Sentiment Analysis, also known as opinion mining, involves building a system to collect and categorize opinions about a product. Automated opinion mining often uses machine learning, a type of artificial intelligence (AI), to mine text for sentiment. Opinion Mining and Sentiment Analysis is a Natural Language Processing (NLP) technique that automatically extracts the opinion, sentiments, attitude, emotions, views etc., in proper context and classify these into different categories like positive, negative, neutral etc. Other terms used for this research domain are subjectivity analysis, subjectivity detection, appraisal extraction and review mining, sentiment mining. The two important tasks involved in Opinion Mining and Sentiment Analysis are (1) Opinion Extraction: extracting the opinionated phrases, in proper context, from free text and (2) Sentiment classification: classifying opinionated phrases based on sentiment orientation.

Sentiment analysis aims to determine the attitude of a speaker, writer, or other subject with respect to some topic or the overall contextual polarity or emotional reaction to a document, interaction, or event. The attitude may be a judgment or evaluation (see appraisal theory), affective state (that is to say, the emotional state of the author or speaker), or the intended emotional communication (that is to say, the emotional effect intended by the author or interlocutor). Opinion mining can be useful in several ways. It can help marketers evaluate the success of an ad campaign or new product launch, determine which versions of a product or service are popular and identify which demographics like or dislike particular product features. For example, a review on a website might be broadly positive about a digital camera, but be specifically negative about how heavy it is. Being able to identify this kind of information in a systematic way gives the vendor a much clearer picture of public opinion than surveys or focus groups do, because the data is created by the customer. It is an extension of data mining which utilizes natural language processing techniques to extract people's opinion from World Wide Web. The recent trend in internet that encourages users to contribute their opinion and suggestion created a huge collection of valuable information in the web. The Opinion mining system analyze each text and see which part contain opinionated word, which is being opinionated and who has written the opinion. Sentiment analysis analyzes each opinionated word or phrase and determines its sentiment polarity orientation, whether it is positive or negative or neutral. It gives the summarized opinion of a writer or speaker.

There are three main classification levels in SA: document-level, sentence-level, and aspect-level SA. Document-level SA aims to classify an opinion document as expressing a positive or negative opinion or sentiment. It considers the whole document a basic information unit (talking about one topic). Sentence-level SA aims to classify sentiment expressed in each sentence. The first step is to identify whether the sentence is subjective or objective. If the sentence is subjective, Sentence-level SA will determine whether the sentence expresses positive or negative opinions. However, there is no fundamental difference between document and sentence level classifications because sentences are just short documents.

Sentiment Analysis Techniques

Machine Learning

Machine learning based Sentiment Analysis or classification can be done in two ways:

1) Sentiment Analysis by using supervised machine learning techniques and
2) Sentiment Analysis by using unsupervised machine learning techniques.

Supervised Machine Learning

In Supervised Machine learning techniques, two types of data sets are required: training data set and test data set. An automatic classifier learns the classification factors of the document from the training set and the accuracy in classification can be evaluated using the test set. Various machine learning algorithms are available that can be used very well to classify the documents. The machine learning algorithms like Support Vector Machine (SVM), Naive Bayes (NB) and maximum entropy (ME) are used successfully in many research and they performed well in the sentiment classification.

Unsupervised Machine Learning

Lexicon Based Method is an Unsupervised Learning approach since it does not require prior training data sets. It is a semantic orientation approach to opinion mining in which sentiment polarity of features present in the given document are determined by comparing these features with semantic lexicons. Semantic lexicon contains lists of words whose sentiment orientation is determined already. It classifies the document by aggregating the sentiment orientation of all opinion words present in the document, documents with more positive word lexicons is classified as positive document and the documents with more negative word lexicons is classified as negative document.

Hybrid Technique

Some researchers combined the supervised machine learning and lexicon based approaches together to improve sentiment classification performance. They considered both general purpose lexicon and domain specific lexicon for determining polarity orientation of sentiment words and feed these lexicons into supervised learning algorithm, SVM. They found that general purpose lexicon performed very poor while domain specific lexicon performed very well. The system classified the sentiment in two steps: First the classifier is trained to predict the aspects and In Next the classifier is trained to predict the sentiments related to the aspects collected in step1.Their system yielded around 66.8% accuracy.

2. Review of status of Research and Development in thesubject

"Naive Bayes and Unsupervised Artificial Neural Nets for Caneun Tourism Social Media Data Analysis" - William B.Claster, Hung Dinh, Malcolm Cooper

Sentiment mining aims at extracting features on which users express their opinions in order to determine the user's sentiment towards the query object. We mine over 70 million Twitter micro-blogs to gain knowledge regarding tourist sentiment on the travel resort destination Cancun in the Yucatan Peninsula of Mexico. We measure sentiment using a binary choice keyword algorithm and a multi-knowledge based approach is proposed using, Self- Organizing Maps and tourism domain knowledge in order to model sentiment. We develop a visual model to express this taxonomy of sentiment vocabulary and then apply this model to maximums and minimums in the time sentiment data. The results show practical knowledge can be extracted.

"Sentiment Classification for Indonesian Message in Social Media" - Aqsath Rasyid Naradhipa, Ayu Purwarianti.

Nowadays, classifying sentiment from social media has been a strategic thing since people can express their feeling about something in an easy way and short text. Mining opinion from social media has become important because people are usually honest with their feeling on something. In our research, we tried to identify the problems of classifying sentiment from Indonesian social media. We identified that people tend to express their opinion in text while the emoticon is rarely used and sometimes misleading. We also identified that the Indonesian social media opinion can be classified not only to positive, negative, neutral and question but also to a special mix case between negative and question type.

Basically, there are two levels of problem: word level and sentence level. Word level problems include the usage of punctuation mark, the number usage to replace letter, misspelled word and the usage of nonstandard abbreviation. In sentence level, the problem is related with the sentiment type such as mentioned before. In our research, we built a sentiment classification system which includes several steps such as text pre-processing, feature extraction, and classification. The text pre-processing aims to transform the informal text into formal text. The word formalization method in that we use is the deletion of punctuation mark, the tokenization, conversion of number to letter, the reduction of repetition letter, and using corpus with Levensthein to formalize abbreviation. The sentence formalization method that we use is negation handling, sentiment relative, and affixes handling. Rule-based, SVM and Maximum Entropy are used as the classification algorithms with features of count of positive, negative, and question word in sentence and bigram. From our experimental result, the best classification method is SVM that yields 83.5% accuracy.

"Scalable sentiment classification for Big Data analysis using Naïve Bayes Classifier" - Bingwei Liu; Erik Blasch; Yu Chen; Dan Shen; Genshe Chen

A typical method to obtain valuable information is to extract the sentiment or opinion from a message. Machine learning technologies are widely used in sentiment classification because of their ability to "learn" from the training dataset to predict or support decision making with relatively high accuracy. However, when the dataset is large, some algorithms might not scale up well. In this paper, we aim to evaluate the scalability of Naive Bayes classifier (NBC) in large datasets. Instead of using a standard library (e.g., Mahout), we implemented NBC to achieve fine-grain control of the analysis procedure. A Big Data analyzing system is also design for this study. The result is encouraging in that the accuracy of NBC is improved and approaches 82% when the dataset size increases. We have demonstrated that NBC is able to scale up to analyze the sentiment of millions movie reviews with increasing throughput.

"Sentiment Analysis for Various SNS Media Using Naive Bayes Classifier and Its Application to Flaming Detection" - Shun Yoshida; Jun Kitazono; Seiichi Ozawa; Takahiro Sugawara; Tatsuya Haga; Shogo Nakamura

SNS is one of the most effective communication tools and it has brought about drastic changes in our lives. Recently, however, a phenomenon called flaming or backlash becomes an imminent problem to private companies. A flaming incident is usually triggered

by thoughtless comments/actions on SNS, and it sometimes ends up damaging to the company's reputation seriously. In this paper, in order to prevent such unexpected damage to the company's reputation, we propose a new approach to sentiment analysis using a Naive Bayes classifier, in which the features of tweets/comments are selected based on entropy-based criteria and an empirical rule to capture negative expressions. In addition, we propose a semi-supervised learning approach to relabeling noisy training data, which come from various SNS media such as Twitter, Facebook, blogs and a Japanese textboard called '2-channel'. In the experiments, we use four data sets of users' comments, which were posted to different SNS media of private companies. The experimental results show that the proposed Na¨ıve Bayes classifier model has good performance for different SNS media, and a semi supervised learning effectively works for the data consisting of long comments. In addition, the proposed method is applied to detect flaming incidents, and we show that it is successfully detected.

"Enhanced Twitter Sentiment Analysis by Using Feature Selection and Combination" - Ang Yang; Jun Zhang; Lei Pan; Yang Xiang

Tweet sentiment analysis is an important research topic. An accurate and timely analysis report could give good indications on the general public's opinions. After reviewing the current research, we identify the need of effective and efficient methods to conduct tweet sentiment analysis. This work aims to achieve a high level of performance for classifying tweets with sentiment information. We propose a feasible solution which improves the level of accuracy with good time efficiency. Specifically, we develop a novel feature combination scheme which utilizes the sentiment lexicons and the extracted tweet unigrams of high information gain. We evaluate the performance of six popular machine learning classifiers among which the Naïve Bayes Multinomial (NBM) classifier achieves the accuracy rate of 84.60% and takes a few minutes to complete classifying thousands of tweets.

"Jakarta Congestion Mapping and Classification From Twitter Data Extraction Using Tokenization And Naïve Bayes Classifier" - Gigih Rezki Septianto; Firman Fakhri Mukti; Muhammad Nasrun; Alfian Akbar Gozali

The data potential of Twitter is a powerful resource for data mining exploration. This research aims to pull the traffic information in Jakarta from Twitter. The first output is to develop a web application that can display Jakarta's traffic situation in real time. The process includes filtering and tokenizing to get the traffic jam's location and direction to be displayed on Google Map. The second output is to develop a predictive analysis system to oversee

Jakarta traffic pattern in a certain period of time using Naïve Bayes Classifier

"A Novel Classifier Based on Meaning for Text Classification" - Murat Can Ganiz; Melike Tutkan; Selim Akyokuş

Text classification is one of the key methods used in text mining. Generally, traditional classification algorithms from machine learning field are used in text classification. These algorithms are primarily designed for structured data. In this paper, we propose a new classifier for textual data, called Supervised Meaning Classifier (SMC). The new SMC classifier uses meaning measure, which is based on Helmholtz principle from Gestalt Theory. In SMC, meaningfulness of terms in the context of classes are calculated and used for classification of a document. Experiment results show that new SMC classifier outperforms traditional classifiers of Multinomial Naïve Bayes (MNB) and Support Vector Machine (SVM) especially when the training data limited.

"Sentiment Analysis and Text Summarization of Online Reviews: A Survey" - Pankaj Gupta; Ritu Tiwari; Nirmal Robert

Sentiment analysis and text summarization has evoked the interest of many scientists and researchers in last few years, since the textual data has become useful for many real-world applications and problems. Sentiment analysis is a machine learning approach in which machine learns and analyze the sentiments, emotions etc. about some text data like reviews about movies or products. These reviews are increasing day by day, due to which summarization of reviews comes in role where summarized form of text in needed, which provides useful information from the large number of reviews. It is very difficult for a human being to extract useful data or summarize it from the very large document. In Text summarization, importance of sentences is decided based on linguistic features of sentences. This paper provides the comprehensive overview of recent and past research on sentiment analysis and text summarization and provides excellent research queries and approaches for future aspects.

"A Brief Review on Sentiment Analysis" - Shreya Banker and Rupal Patel

To mine the opinion on the web, it is essential to perform a well-defined task, which helps us to retrieve the information from the available data on the web. We have started this discussion with the introduction on sentiment analysis, which gives us an insight into sentiment analysis. The detail discussion on various methods proposed by different researchers is also presented. Different types of sentiment analysis techniques give a research

direction in different directions. Finally, a method is proposed based on the Naïve Bayes classifier.

International Status: NIL

National Status: NIL

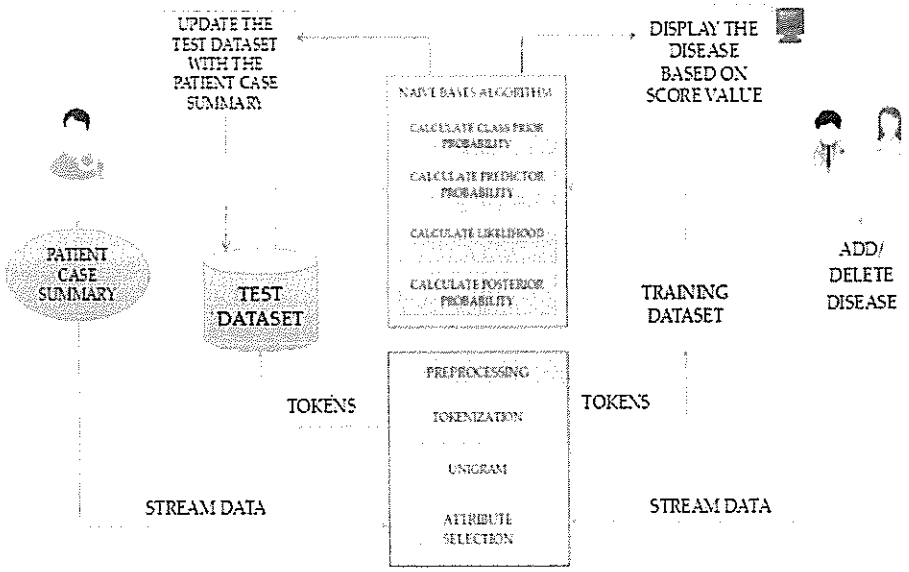
3. Progress/achievement so far,
   a) Reference papers was collected.
   b) Literature survey was studied.
   c) Proposal work has been started with hybrid data mining techniques in identifying diseases of the patients. This system can be used by all patients or their family members who need help in emergency.

4. **Work Plan:**

4.1 Methodology:

- In Proposed System, we are applying hybrid data mining techniques in identifying diseases of the patients. This system can be used by all patients or their family members who need help in emergency.
- To Apply hybrid data mining techniques to the disease diagnosis benchmark data set to establish baseline accuracy for each single data mining technique in the diagnosis of disease in patients.
- To Apply the same hybrid data mining techniques used in disease diagnosis to data set to investigate if single data mining techniques can achieve equivalent (or better) results in identifying suitable treatments as that achieved in the diagnosis.

Architecture Diagram



Module Description

Registration and Login

This module is the initial stage of the system. The first step involves registration. It will get users' information such as name, email id and password as input. This information is stored in MySql database in form of table. After the registration, the users' information is verified by the admin. In the Login module, the information entered by the user is verified with the help of MySql database. If the information given by the users matches with the database, then user can utilize the system otherwise the access is denied.

Preprocessing

Information entered by the user is pre-processed with the help of OpenNLP tool. It performs tokenization by using the Unigram method. The Unigram method will split the data streams into tokens. After performing tokenization, attribute selection is done. It is also called as feature selection or variable subset selection and it is the process of selecting a subset of relevant features (variables, predictors) for use in model construction.

Disease Prediction

This is the important module that involves in classification of the diseases. Naive Bayes Classifier is used in classification of the diseases. After the completion of pre-processing, the tokens are given to the Naive Bayes Classifier. The Classifier performs three steps for classifying diseases, namely, finding the class prior probability, calculating the likelihood, calculating the prior probability of predictor and finally finding the posterior probability of class. After the calculation of posterior probability, the class which has the highest probability

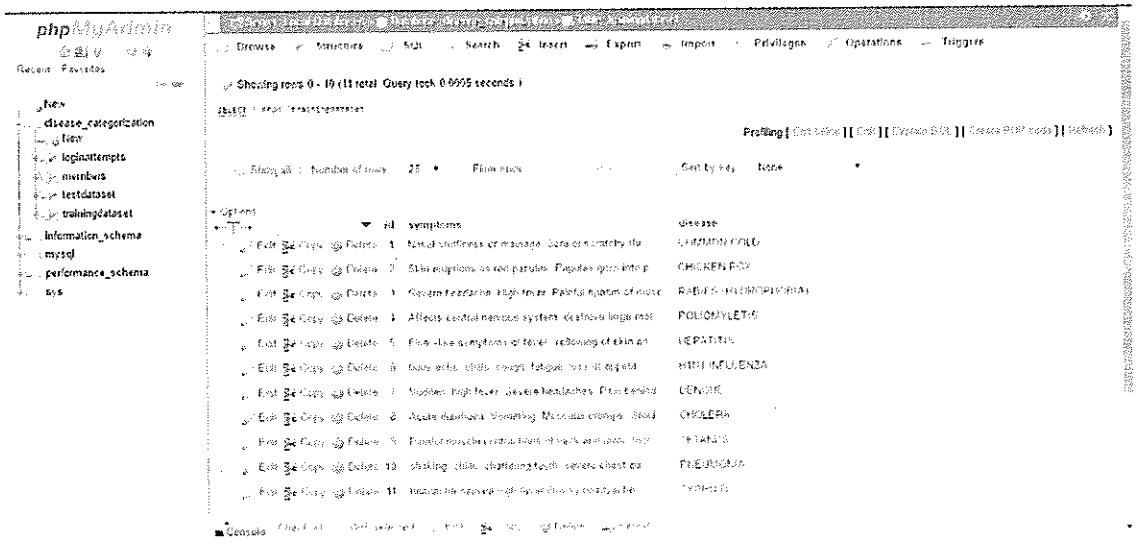
score value is declared as the disease.

Add Disease

In this module, only doctor has the authority to add disease and symptoms related to the disease. After the pre-processing, the entered symptoms and disease is stored in MySql database.

Training Dataset

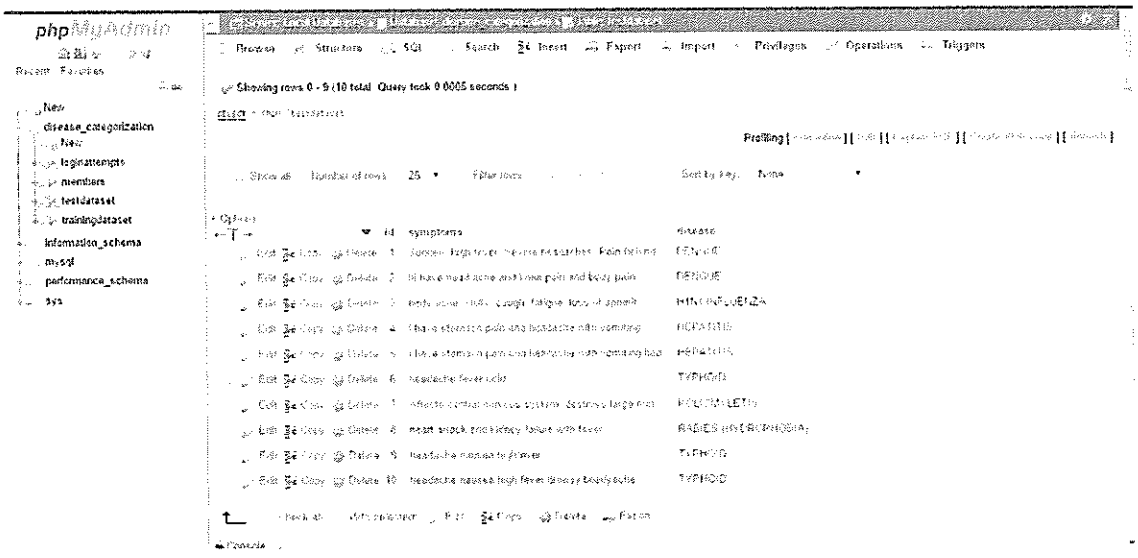
It contains the diseases and symptoms related to those diseases.



Training Dataset

Test Dataset

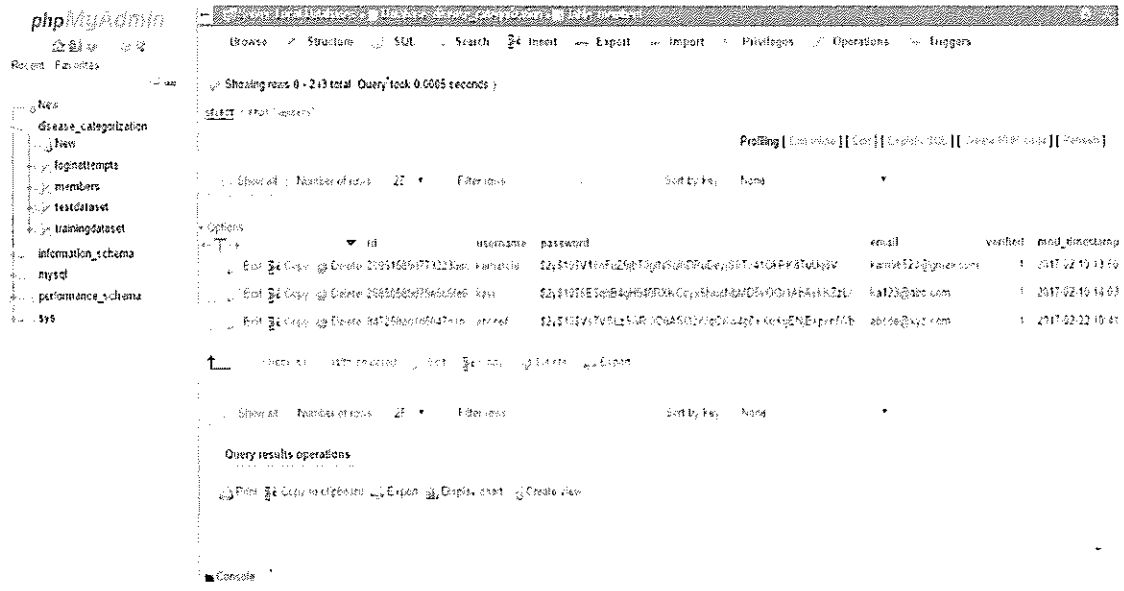
It contains the history of patient case summary and diseases diagnosed.



Test Dataset

Members

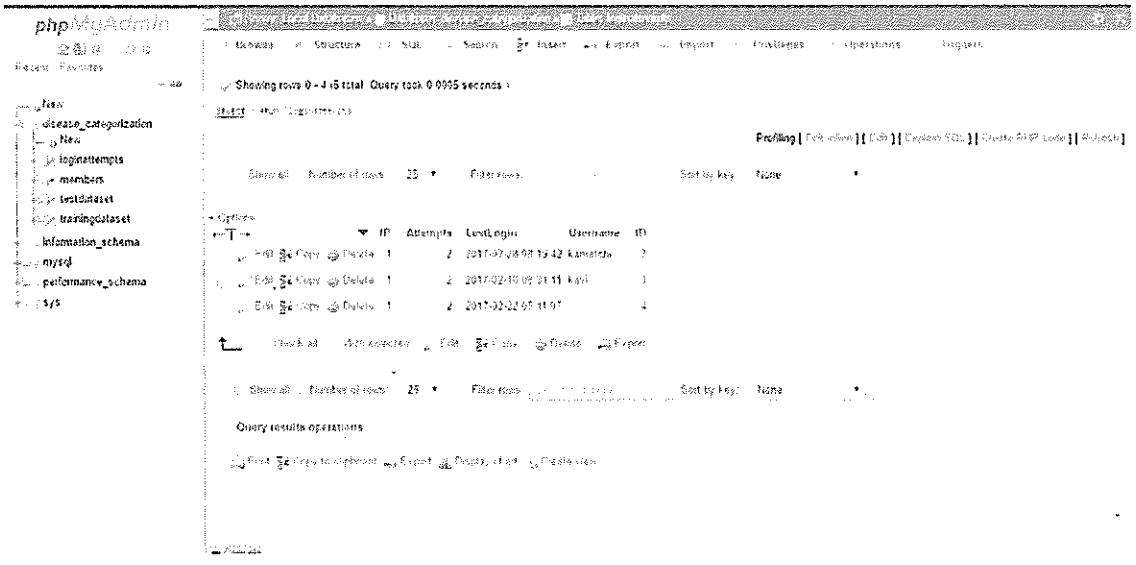
It contains the details of the members registered.



Members

Login Attempts

It contains the login attempt details of the users.

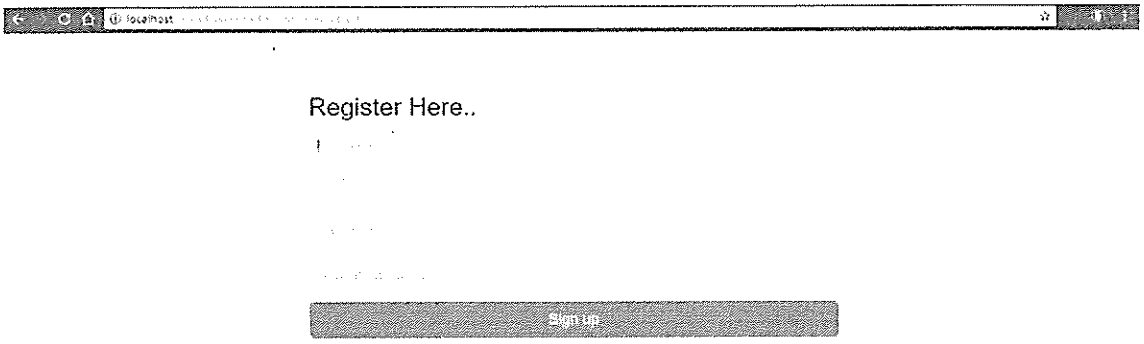


Login Attempts

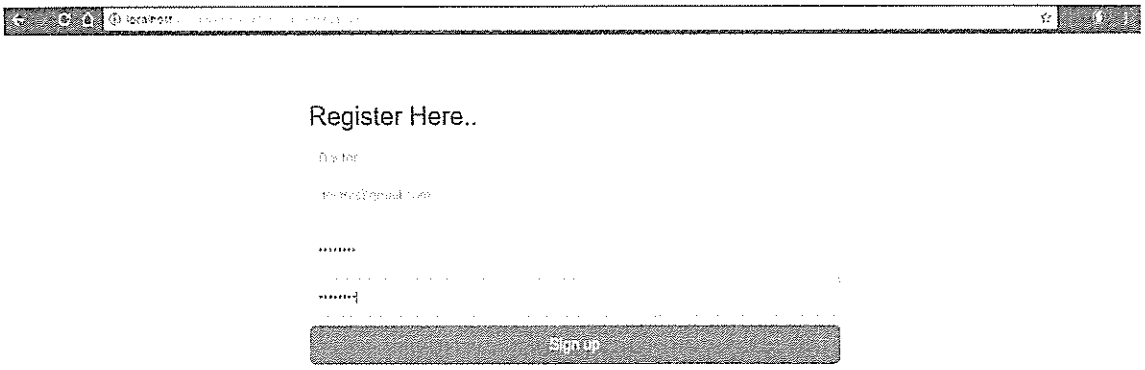
Doctor Activities

Doctor Registration

It allows the doctors to register by entering the details required in the registration form and signing up.

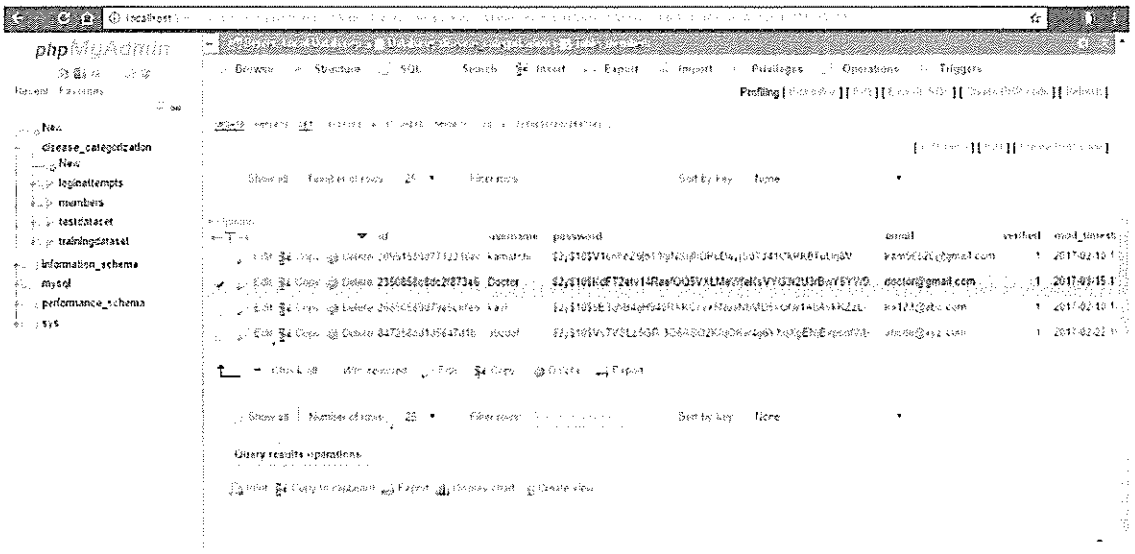


(a) Doctor Registration Form



(b) Doctor Registration

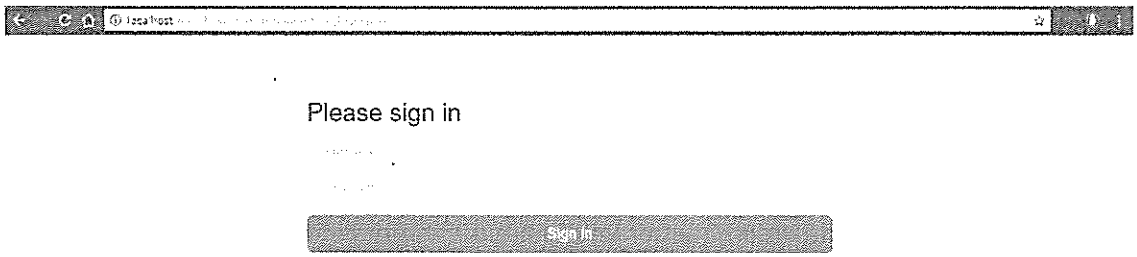
After signing up the details given by the doctor is stored in the members data set.



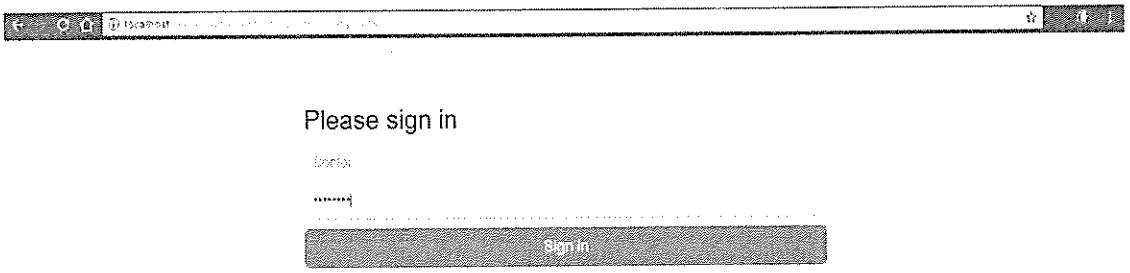
(c) Doctor Registration Information Updated In Members Dataset

Doctor Login

The doctor can login with the registered information.



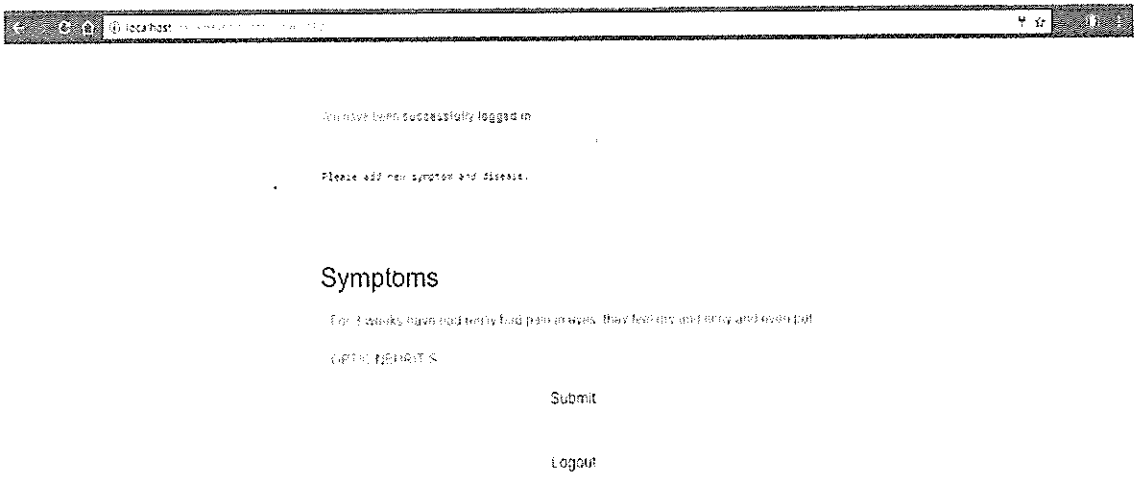
(a) Doctor Login Form



(b) Doctor Login

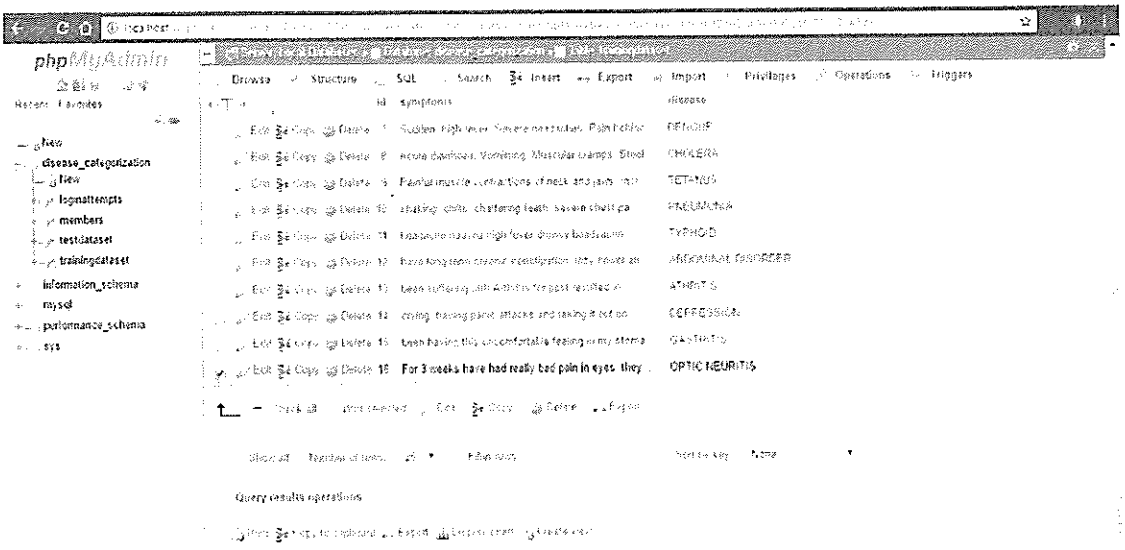
Adding Disease

After logging in, the doctor can add the disease and its related symptoms and submit it.



(a) Adding Disease

The information entered by the doctor is updated in the training data set.

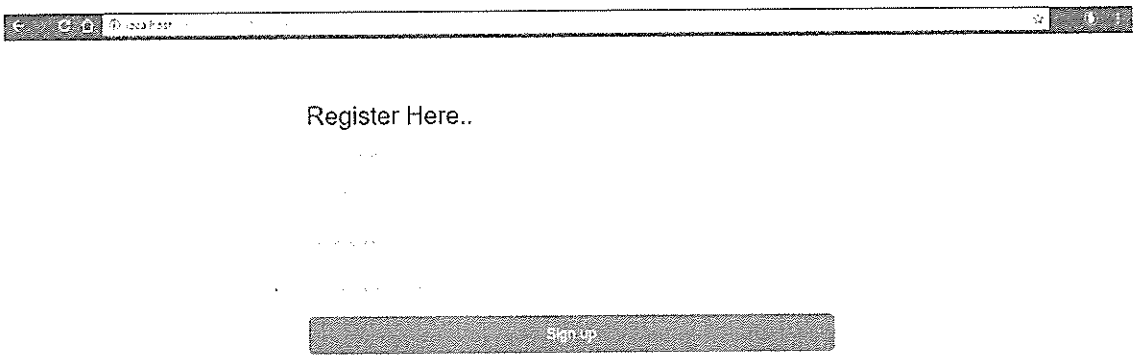


(b) Disease Updated in The Training Dataset

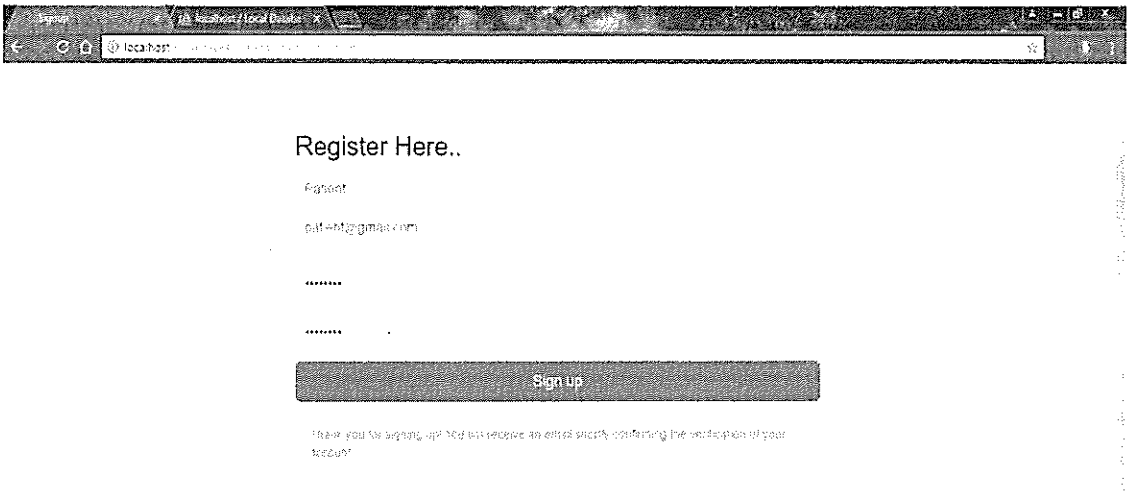
Patient Activities

Patient Registration

It allows the patient to register by entering the details required in the registration form and signing up.



(a) Patient Registration Form



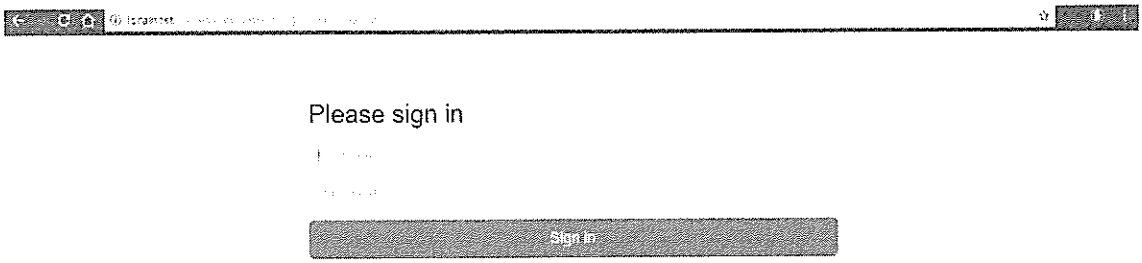
(b) Patient Registration

After signing up the details given by the patient is stored in the members data set.

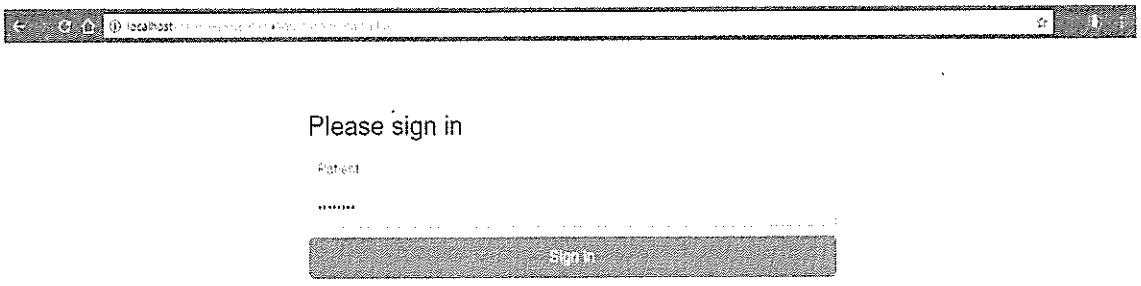
(c) Patient Registration Information Updated In Members Dataset

Patient Login

The patient can login with the registered information.



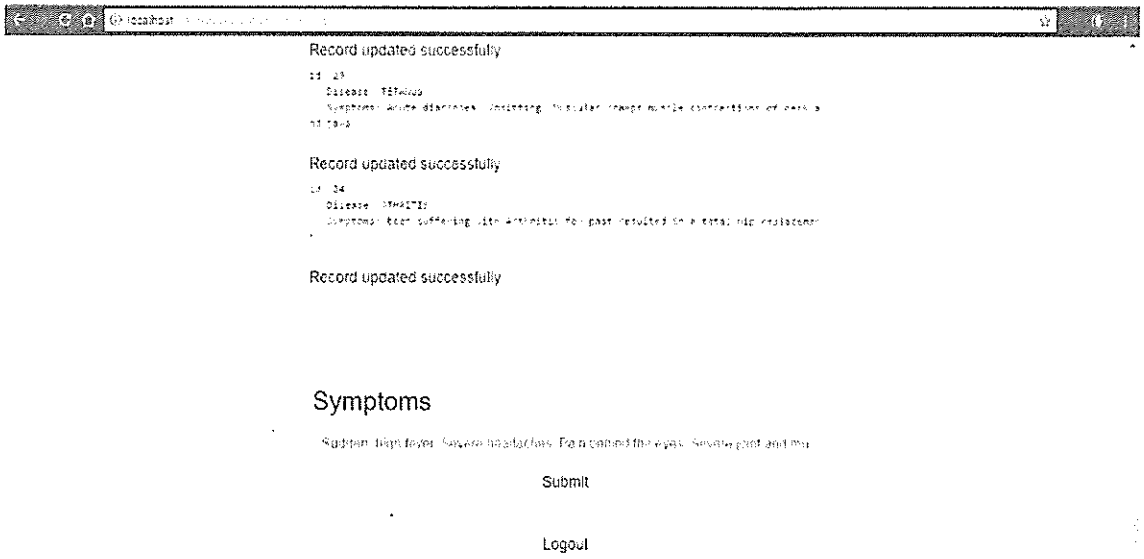
(a) Patient Login Form



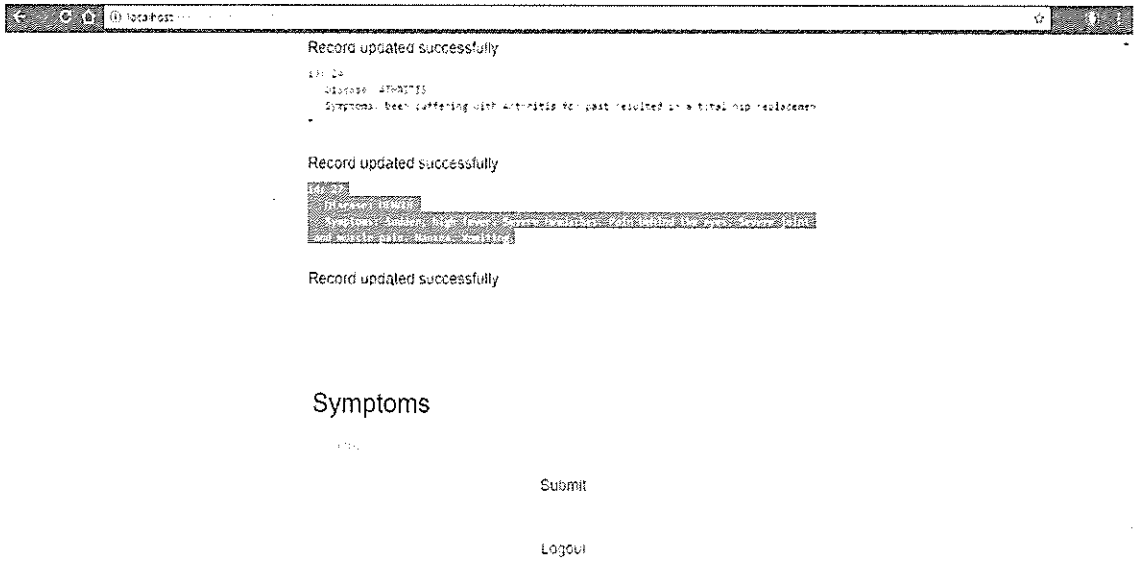
(b) Patient Login

Entering Symptoms

After logging in, the patient can enter their symptoms and submit it.

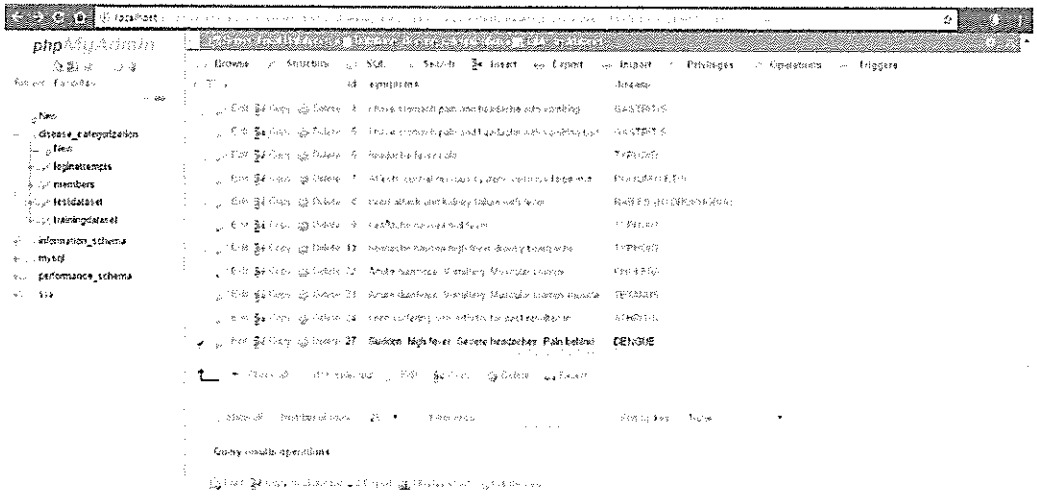


(a) Entering Symptoms The disease related to the entered symptom is displayed.



(b) Displaying Result

After displaying the result, it will be updated in the test data set.

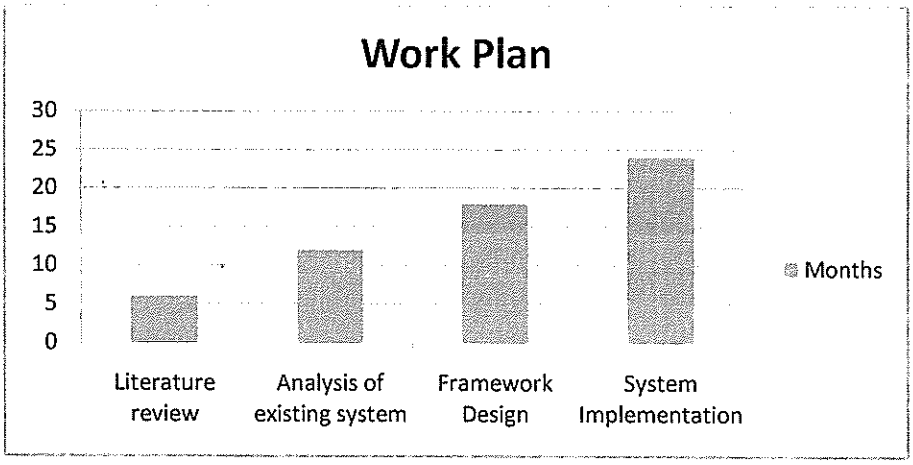


(c) Updated Test Dataset

4.2 Time schedule of Activities giving milestones through BAR diagram

Work plan (including detailed methodology and time schedule)

Sl. No	Activity / Milestone	1 Year		2 Year	
1	Literature review	1-6			
2	Analysis of existing system		7-12		
3	Framework Design			13-18	
4	System Implementation				19-24



4.3 Expected outcome within the time period of Seed Money Scheme

It is used to identify the diseases using Classification Algorithms.

5. Suggested Plan of action stating the name of funding agency where the project will be communicated for financial support within the time period of project.

NIL

6. Bibliography: NIL

7. List of Projects submitted/implemented by the Investigators (Separate for Pi and Co-PI)

NIL

7.1 Details of Projects submitted to various funding agencies:

Sl. No.	Title	Cost in lakhs	Month of submission	Role as PI/ Co-PI	Agency	Status
NA						

7.2 Details of Projects under implementation

Sl. No.	Title	Cost in lakhs	Duration	Role as PI/ Co-PI	Agency
NA					

7.3 Details of Projects completed during the last 5 years

Sl. No.	Title	Cost in lakhs	Duration	Role as PI/ Co-PI	Agency
NA					

8. List of publications published by the Investigators, if any:

Principal Investigator

S.No	Author Names	Title of the paper	Name of the Journal	Vol (Issue)	Page No	Year
1.	Dr.S. Thiru Nirai Senthil and Dr.A. Muthukumaravel	Enhanced Secure Cloud Storage Auditing System with Proxy and BLS Signature Algorithm	Journal of Advanced Reseach in Dyanamical and control Systems	11(1)	522-527	2019
2.	S. Kannan and Dr.S.Thiru Nirai Senthil	Fuzzy based Smart Greenhouse Hydroponic Control System Using IOT and Cloud Technology	Journal of Advanced Reseach in Dyanamical and control Systems	11(5-Special Issue),	2320-2327	2019
3.	Dr.S. Thiru Nirai Senthil, S. Kannan and L. Lalli Rani	An Secure Ordered Bucketization in Network Security in Dealock Starvation Technology	Journal of Advanced Reseach in Dyanamical and control Systems	10(3)	442-450	2018
4.	N.Mathimagal, Dr.S.Thiru nirai Senthil and Dr.A. Muthukumaravel	Social Network Analytics (SNA) Fraud	*International Journal of Pure and Applied Mathematics*	118(20)	191-202	2018
5.	G. Preethi, Dr.S.Thiru nirai Senthil and Dr.A. Muthukumaravel	End-To-End Per-Packet Delay In Multi-Hop Wireless Ad-Hoc Networks Tomography	*International Journal of Pure and Applied Mathematics*	118(20)	245-253	2018

Co- Principal Investigator

S.No	Author Names	Title of the paper	Name of the Journal	Vol (Issue)	Page No	Year
1.	S. Kannan and Dr.S.Thiru Nirai Senthil	Fuzzy based Smart Greenhouse Hydroponic Control System Using IOT and Cloud Technology	Journal of Advanced Reseach in Dyanamical and control Systems	11(5-Special Issue),	2320-2327	2019
2.	Dr.S. Thiru Nirai Senthil, S. Kannan and L. Lalli Rani	An Secure Ordered Bucketization in Network Security in Dealock Starvation Technology	Journal of Advanced Reseach in Dyanamical and control Systems	10(3)	442-450	2018
3.	Dr.A.Muthuku maravel, S.Kannan, S.Thiru Nirai Senthil.	A High Load Balancing and Enforcing Quality of Service In Manets Using Hybrid Algorithm	International Conference on Recent Trends in Science & Management			2018
4.	S.Thiru Nirai Senthil., S.Kannan., Dr.A.Muthuku maravel.	A High Performance UDP-Based Data Transfer For High – Speed Wide Area Networks	International Conference on Recent Trends in Science & Management			2018

9. Budget*

Sl. No.	Head	Amount in INR
1	Minor Equipment/software (Generic Name with minimum required accessories, make & model & cost in Indian Rupees)(Mention other PI(s) name, Department with whom joint purchase is proposed)	50000
2	Consumables (Like ICs, application boards, chemicals, testing charges, tools, etc.)	25000
3	Travel support for the purpose of research work.	15000
4	Contingency	5000
5	Others	5000
	Total	100000

* In case of any joint proposal for purchasing a same equipment, each of the associated PIs is also required to give separate budget(without any clubbing) to avoid any ambiguity, if all the associated projects are not awarded by committee.

10. Name of at least two subject experts from the Institute and one from the outside Institute with their contact details:

a) Dr. A.Muthukumaravel, Dean, Faculty of Arts and Science, BIHER , Chennai-73.

b) Dr. T.Kamalakannan, Head, Dept of BCA&IT, VISTAS, Chennai-117.

CERTIFICATE FROM THE INVESTIGATOR

Project Title: Prediction of Various Diseases using Sentimental Analysis

It is certified that

1. I do hereby agree to submit a complete proposal for financial support to the external funding agency within the time period of SMS-2018
2. I undertake that spare time on equipment procured in the project will be made available to other users.
3. I agree to submit a certificate from Institutional Biosafety Committee, if the project involves the utilization of genetically engineered organisms. I also declare that while conducting experiments, the Biosafety Guidelines of Department of Biotechnology, Department of Health Research, GOI would be followed in to.
4. I agree to submit ethical clearance certificate from the concerned ethical trails/experiments/exchange of committee, if the project involves field specimens, human & animal materials etc.
5. I agree to abide by the terms and conditions of SMS-2018, BIHER, and Chennai.

Name and Signature
Principal Investigator

Name and Signature
Co-Principal Investigator

Date:05.10.2019

Place: Chennai

Forwarded by the Head of the Department

Signature of the Head

PROJECT EVALUATION FORMAT

Recommentation Sheet

Name of the Principal Investigator	Dr.S.ThiruNirai Senthil
Name of the Co-Investigator	S.Kannan
Name of the Department	Computer Application
Title of project	Prediction of Various Diseases using Sentimental Analysis
Recommendation of the evaluation committee	Recommended
Financial allocation recommended	Rs. 1 lac (One lakh only)

Sl. No.	Head	Amount in INR
1	Minor Equipment/software (Generic Name with minimum required accessories, make & model & cost in Indian Rupees)(Mention other PI(s) name, Department with whom joint purchase is proposed)	50000
2	Consumables (Like ICs, application boards, chemicals, testing charges, tools, etc.)	25000
3	Travel support for the purpose of research work.	15000
4	Contingency	5000
5	Others	5000
	Total	100000

Name and Signature of the Research Advisory Committee members with date

(Dr. P. Naveenchandran)

Internal Quality Assurance Cell IQAC Selaiyur Chennai-79. BIHER