

Journal of
Information Technology

Volume 6, Issue 1, 2018-19

JIT

Volume 6

Issue 1

2018-19



Department of Information Technology

Pillai College of Engineering

Plot No. 10, Sector 16, New Panvel - 410206

Maharashtra, India.



Journal of Information Technology (JIT)

JIT, Volume 6, Issue 1 2018-19

Editor-in-Chief

Dr. Satishkumar L. Varma

Editorial Board Members

Dr. Satishkumar L. Varma

Prof. Sushopti Gawade

Prof. Gayatri Hegde

Prof. Madhu Nashipudimath

Message



Dr. Sandeep M. Joshi
Principal, PCE

Good health is a prerequisite to human productivity and development. Healthy efforts are being made to bring Department Journals, which is an outcome of the research work carried out by students and faculty.

The journal will definitely help to showcase the research activities that are happening in the campus. It also helps in building up teamwork which is very much needed today in the world of competition.

I encourage and appreciate the efforts of the team of Journal of Information Technology and convey all my best wishes to them.

Editorial



Dr. Satishkumar L Varma
Editor-in -Chief

Dear faculty and students of Pillai College of Engineerig,
Greetings!

It is with deep satisfaction that I applaud and congratulate you for contributing technical papers. I feel proud to bring out this issue of the Journal of Information Technology (JIT).

This journal focuses on a variety of topics such as Machine Learning, IoT, Wireless Networks and Android Development. This Issue of JIT explores in detail, the application of Machine Learning to the problems such as employability prediction, heartbeat classification, terror activities detection, sentiment detection and classifying documents. It also discusses in detail the application of IoT to create smart locks, energy consumption and safety of children and women.

This issue covers twelve papers published by faculty and under-graduate students of Department of Information Technology, Pillai College of Engineering (PCE). I am happy to note that this issue of PCE JIT will be helpful for the future engineers working in the areas of IoT, Machine Learning and Wireless Sensor Networks.

I would like to extend my best wishes to all our students and teachers for contriuting to this issue.

We are honored to dedicate the issue of JIT to all the students and faculty of PCE.

Contents

CV Based Employability Prediction using Machine Learning	1-5
Nithik Pradeep, Alwin Alex, Alam Ansari, Manivel Chettyar, Shubhangi Chavan.	
ECG Based Heartbeat Classification	6-16
Ashwin Nair, Nikhil Elayath, Shashank Nair, Hariharan V.	
Employee Tracking and Attendance Management System Using RFID	17-21
Priya Rajput, Sofiya Rao, Laxmi Tanavarappu, Manali Thombare, Krishnendu Nair.	
IoT Based Energy Monitoring System	22-27
Sameer Mhatre, Vishal Patil, Jaiprakash Khichi, Chandan Chodhary, K.S.Charumathi.	
IoT Based Women And Children Safety System	28-32
Shrayesh Kanade, Siddhi Morajkar, Priyanka Borse, Vrutant Mehta, Rupali Nikhare.	
Literature Survey on Real Time OSN Analysis to Detect Online Terrorists	33-38
Mrudul Bornare, Tasneem Attarwala, Riya Jadhav.	
Mobile App for Stress Detection and Mental Health	39-43
Shruti Pawar, Amruta Salvi, Shifa Khan, Neha Gawand, Madhumita Chatterjee.	
Sentiment Analysis Based on Comments from Online Social Network	44-46
Vedant Patil, Jayesh Thakur, Kapildev Yadav, Deepti Lawand.	
Text Document Clustering using Latent Semantics Indexing	47-51
Ankita More, Ameya Pokharkar, Aditya Sawant, Mayur Walshinge, Madhu Nashipudimath.	
Chatbot Based Question Answering System	52-56
Giridhar Srinivasan, Voval Jain, Prashant Niladhe, Vishal Gupta, Deven kanse, Shubhangi Chavan.	
Real Time Traffic Event Detection using Tweet Stream	57-60
Mohit A Rai, Sumod Menon, Riya Sawant, Devbrat Singh, Sagar Kulkarni.	

Contents

IOT Based Portable Smart Lock	61-65
Ayush Shetty, Manthan Parvadia, Onkar Pokharkar, Shubham Shinde, Payel Thakur.	

About the Editors

Satishkumar L. Varma received his Ph.D degree in Computer Science and Engineering under the guidance of Dr. S N Talbar from SGGS I E & T, SRTMU, Nanded, India in March 2013. He received his graduation and postgraduation degree in Computer Engineering from Dr. BATU, Lonere, Raigad, MH, India, in the year 2000 and 2004, respectively. He is currently working as Professor and Head in the Department of Information Technology, Pillai College of Engineering, New Panvel, MH, India. He has twenty-one years of experience in teaching and research. He has received and successfully executed three R&D Funded Projects of amount more than Rs 9 Lakhs. He has published 1 copyrights, 8 Book Chapters, more than 32 refereed Journal papers and more than 36 papers in referred National as well as International Conferences including IEEE, Springer and IET with a second best paper award at National level paper presentation competition in Threshold-2000. He is recognized as Teacher of University of Mumbai in Ph.D Degree in Computer Engineering and Information Technology. His delivered talks include Image Processing, Object Oriented Analysis and Design, MATLAB, Scilab, Hadoop, LaTeX, Android, Python, R, Google Scripts and Docs. He is a member of Technical Professional society in IEEE, ISTE, and CSI. His research interests involve Digital Image and Video Processing, Medical Imaging, AI and Machine Learning, Soft Computing, Data Mining and Information Retrieval.

Sushopti Gawade is pursuing Ph.D in Computer Engineering with research area Usability Engineering in Agriculture Domain. She has received B.E in Computer Science and Engineering in 1997 and M.E Computer Science and Engineering from Walchand College of Engineering Sangli in 2006. Currently she is working as a Professor in Pillai College of Engineering, Panvel. She is highly dedicated and performance-driven professional with 21 years of teaching experience in Mumbai University. She has ability to coordinate and direct all phases of project-based efforts while managing, motivating, and leading the project team. She is an excellent problem solver and opportunities identifier to improve and resolve critical issues. She is quick learner of new concepts and technologies and has excellent ability in expressing ideas clearly and good team management skills.

Gayatri Hegde is pursuing Ph.D degree in Computer Engineering from Thadomal Shahani Engineering College, University of Mumbai. She has received her M.E in Computer Engineering from Pillai College of Engineering, Mumbai University. She has received M.B.A degree in Systems and Marketing from Sikkim Manipal University and completed B.E in Computer Science and Engineering from Basaveshwar Engineering College, University, Karnataka. She is currently working as assistant professor in Pillai College of Engineering, New Panvel, Maharashtra since 2010. She has 7 conference and journal publications and has attended 5 FDP. Her area of interest includes Operating system, Cloud Computing, Big Data Analytics and Distributed Systems.

Madhu Nashipudimath is pursuing her Ph.D. from P.A.H.E.R University, Udaipur in the field of big data analytics. She has received her B.E degree in Computer Engineering from B.L.D.E.A's V.P. Dr.P. G.Halakatti College of Engineering and Technology affiliated to VTU of Karnataka. She has completed her post graduation in Computer Science from Walchand College of Engineering Sangli affiliated to Shivaji University, Kolhapur Maharashtra. She has more than 25 years of teaching experience. She is presently working as Assistant Professor in department of Information Technology of Pillai College of Engineering Navi Mumbai. She has published about 31 papers in International and National Journals and Conferences and also attended 32 conferences, workshops, FDP and training programmes. She is a recognised Teacher of the University of Mumbai for P.G. Programme in IT and is actively involved in University activities like designing syllabus revision for UG and PG programmes. She is a reviewer of reputed journals like International Journal Big Data and Springer. She had submitted her research thesis titled “ Novel approaches of Integration and indexing in Social Media for Big Data Analysis”. Her fields of interest are Data Mining, Big data, Information Retrieval and Fuzzy logic.

CV BASED EMPLOYABILITY PREDICTION USING MACHINE LEARNING

Nithik Pradeep, Alwin Alex, Alam Ansari, Manivel Chettyar, Prof. Shubhangi Chavan

Department of Information Technology, PCE, Navi Mumbai, India, 410206

nithikpradeep12@gmail.com

alwinalex9797@gmail.com

ansariehtesham.ae@gmail.com

manivelchettyar@gmail.com

srathod@mes.ac.in

Abstract—In the last few years, the number of job opportunities across the country has increased with the advancement of different technologies across different fields. But the rate of rejection of various candidates has also risen. In this work, we develop a set of techniques that make the recruitment process effective and transparent. We will develop a system that will take the CV of each candidate as an input and then based on his skills, experience, academics, etc mentioned in the CV will rank him in comparison with the other candidates and determine whether or not he/she is eligible for the job. Machine Learning plays an important part in this work as the important modules like feature extraction and ranking will be done using machine learning algorithms like k nearest neighbor and support vector machines (svm) .

Keywords: Employability, Machine Learning, K - Nearest Neighbor, Support Vector Machine.

I. INTRODUCTION

Employability is a set of achievements, understandings and personal attributes that make individuals more likely to gain employment and to be successful in their chosen occupations. It is the ability of the candidate to check whether he is capable to gain employment or not. Also, from a recruiting companies point of view, it is a very tedious job to go through each and every resume manually and then sort the candidates according to their requirements. It would be largely beneficial if this work could be done in a way that would save time and manual efforts.

CV based Employability Prediction using Machine Learning, refers to the use of different algorithms to predict the employability and the ranking of the candidates in comparison to other candidates. The job of recruitment is mainly done manually by recruiters and this system reduces that manual effort to a great extent.

The basic idea of this project came from after the referral of various research papers and understanding the technology used in past project research. The various techniques that have been used in these research papers include machine learning algorithms like K - Nearest Neighbor, Naive Bayes, Decision Tree, Support Vector Machine, etc. In this chapter relevant techniques in literature is reviewed. It describes various techniques used in the work.

II. SCOPE OF THE PROJECT

The software we are implementing will be able to parse and rank resumes as per the information mentioned within the resumes for the applying candidates. The software will use machine learning algorithms and rank the resumes as per the final results. The candidates can give their cv's as input in the pdf or .doc format and the system can rank any number of candidates. This will reduce the manual work that the recruiters have to do and ensure the candidates of fair selection.

III. RELATED WORK

Large number of research paper and information related to this cv analysis system is present online and was reviewed for this work.

Neeraj Khadilkar, Deepali Joshi proposed, “ Predictive Model on Employability of Applicants and Job Hopping using Machine Learning” [2]. Here, in this paper they are trying to predict whether a candidate is employable or not and also whether the candidate will leave the job after a specific amount of time which is called Job Hopping.

Text mining and appropriate weighing is used for screening the resumes along with the personal information of candidates [2]. Among the several algorithms used, they got the best accuracy for naive bayes for employability prediction. They have accepted the input resumes in only one format which is a restriction for the candidates.

G.Vadivu, K.Sornalakshmi proposed, “Applying Machine Learning Algorithm for Student Employability Prediction using R” [3]. Here, in this paper they have applied the machine learning algorithms kNN and naïve Bayes to predict the future performance which will be useful for the students to improve themselves to get placement through campus.

The categories considered for predicting the student employability are basic concept, programming skills, mathematics, advanced technical skills, etc. Extra co-curricular activities has not been included in this data which

can be included in the future works. Both the algorithms give the output as Yes/No for employability. The algorithms were applied on the data set of 250 students. The accuracy obtained after analysis for KNN is 95.33% and for the Naïve Bayes is 97.67% [3].

Evanthia Faliagka, Kostas Ramantas, Athanasios Tsakalidis, Giannis Tzimas proposed, “Application of Machine Learning Algorithms to an online Recruitment System” [7]. In this work, the candidate’s LinkedIn profile is used to extract information about the candidate and do the evaluation.

The system ranks the applied candidates using machine learning algorithms like Linear Regression, Regression Tree and Support Vector Regression. But this approach requires sufficient training data as an input, which consist of previous candidate selection decisions [7]. Extroversion personality traits which are crucial for job which involves interaction with customers are extracted from the candidate’s LinkedIn profile and blogs. Although using LinkedIn profiles can be risky as candidates can enter false information.

Rajendra.S Choudhary, Rajul Kukreja, Nitika Jain, Shikha Jain proposed “Personality and Education Mining based Job Advisory System [5]. Here in this work the main fact was MTBI algorithm they mainly focused on personality along with educational qualification.

They have mapped results on five parameters Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism. They called it as OCEAN. The candidate is asked to fill the form provided by the system due to this preprocessing becomes easier, due to this complexity is reduced. This, OCEAN maps value of each for candidate which is the suitable job position.

In 2016, Tripti Mishra ,Dharminder Kumar, Sangeeta Gupta proposed “Students’ employability prediction model through data mining”. In this work they have used various classification algorithm such as Bayesian methods, Multilayer Perceptions and Sequential Minima Optimization (SMO), Ensemble Methods and Decision Trees, to predict the employability.

In this they found that highest accuracy of 70.31% was achieved by random forest algorithm. This algorithm execution was 0.11 sec. Another algorithm J48 took time of just .02 sec and its accuracy was 70.19%.the tool use for classification was WEKA. In the result they compared all the classified techniques used for classification.

IV. INFERENCES

From going through related works on our project and above mentioned research paper we have derived following inferences. Although some good work has been done to create a CV analysis system, but most systems do not provide a great accuracy. Also, the input required in most of these systems is the cv only in pdf format. Moreover in some papers not all the important details about the candidate is considered before

making the prediction.

V. PROPOSED SYSTEM

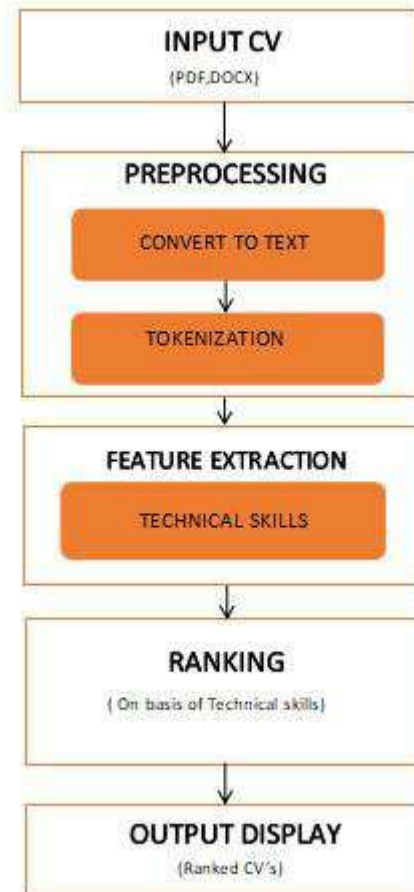


Fig. 1 Employability Prediction System

The Proposed system architecture consists of several modules interacting with each other to accept a CV in pdf format and rank it using machine learning algorithms. The Architecture is shown in Figure [1].

1. CV of the candidate will be taken as input in pdf or doc format.
2. Conversion to text and tokenization is done using PDFMiner.
3. The required criteria can be set and the cv’s can then be classified using K - nearest neighbor algorithm.
4. The cv’s which match the criteria will then be ranked based on their score using Knuth Morris Pratt(KMP) algorithm.
5. The final ranked list of candidates will be displayed.

VI. METHODOLOGY

The system consists of the following modules:-

1. Data Input

The CV's of the candidate should be taken as input in pdf/doc format which will then be tokenized to create a dataset.

2. Pre - Processing

1) *Convert to Text*:- The document file needs to be converted to text file before performing tokenization. Various algorithms can be used to convert pdf to text.

2) *Tokenization*: The process of segmenting running text into words and sentences. Naturally, before any real text processing is to be done, text needs to be segmented into linguistic units such as words, punctuation, numbers, alpha-numeric, etc. This process is called tokenization. Both the above methods can be performed using a tool called PDFMiner.

PDFMiner is a tool for extracting information from PDF documents. It focuses entirely on getting and analyzing text data. PDFMiner allows one to obtain the exact location of text in a page, as well as other information such as fonts or lines. It includes a PDF converter that can transform PDF files into other text formats. It has an extensible PDF parser that can be used for other purposes than text analysis.

Input:- All the CV's taken as input from the candidate or the admin is given to the pdfminer algorithm.

Output:- pdfminer will convert these CV's to text format and then tokenize them into individual lexical units.

3. Feature Extraction

Once we have taken the CV's of the candidates as input and then parsed them, we will extract the features from the CV. The major features that we have considered are: - Academic Performance, Experience, Technical Skills. The recruiters can set the parameters for each attribute and then the system will classify the cv's which match with the criteria. A machine learning algorithm called K - Nearest Neighbor is used to perform this module.

K- Nearest Neighbor is a machine learning algorithm that can be applied to the data from any distributions. The algorithm is very simple and it gives the good classification for large number of samples. An object (new instance) is classified by the majority votes for its neighbor classes. The object is assigned to the most common class. Calculate the distance between new example ϵ and all examples in the training set.

Euclidean distance between two examples.

$$- X = [x_1, x_2, x_3, \dots, x_n] \quad - Y = [y_1, y_2, y_3, \dots, y_n]$$

- The Euclidean distance between X and Y is defined as:

$$D(X, Y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

Input:- The tokenized CV's which was saved in the array by pdfminer is given as input to the KNN algorithm.

Output:- When the admin enters the query, the KNN algorithm will search in all the CV's and find the nearest neighbor to the query entered.

4. Ranking

After feature extraction is done the candidates will be ranked in comparison to each other. Scores will be generated for each candidate that matches the required criteria and using machine learning the rank would be generated. As the score is a continuous variable, the candidate ranking problem can be reduced to a regression problem where the candidate score must be learned using supervised learning techniques. Then the system outputs the final ranked list by applying the learned function to sort the candidates. This can be done using another algorithm called Knuth Morris Pratt(KMP) algorithm.

In computer science, the Knuth–Morris–Pratt searching algorithm (or KMP algorithm) searches for occurrences of a "word" W within a main "text string" S by employing the observation that when a mismatch occurs, the word itself embodies sufficient information to determine where the next match could begin, thus bypassing re-examination of previously matched characters.

Knuth Morris Pratt (KMP) is an algorithm, which checks the characters from left to right. When a pattern has a sub-pattern appears more than one in the sub-pattern, it uses that property to improve the time complexity, also for in the worst case.

Input:- All the CV's which had the searched skill was selected by the KNN algorithm and those selected CV's are given to the KMP algorithm.

Output:-Score of each CV's is obtained as output which is then used to rank the CV's to provide the final output.

5. Result

The result page will display the output of the ranking module. It will show the names and details of all the candidates who match the required criteria with the proper rank.

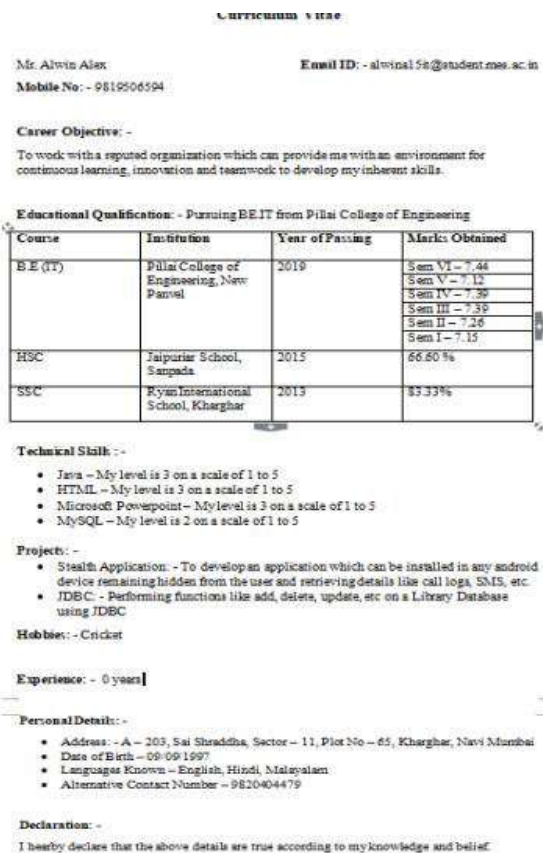


Fig. 2 Sample of Input CV

VII. RESULTS



Fig 3. Skill Searched by Admin for Ranking

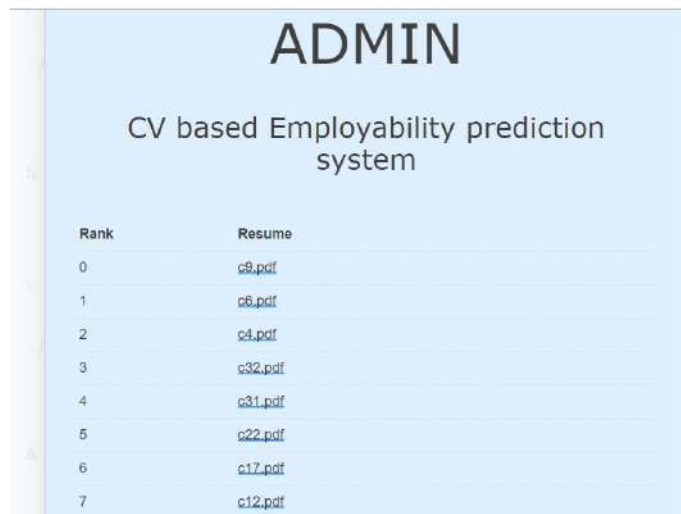


Fig 4. Ranked Result for Skill Python

To check the efficiency of the system at lowest level, simple programming skills as criteria are used. We have analyzed and tested the system for 20 different types of skills. Table 4.2. shows the testing process by specifying the number of skills tested for a particular type along with the number of CV's that are correctly ranked as well as the number of CV's that are wrongly ranked. Thus, by finding the correctly ranked CV's we find the accuracy for the whole system.

TABLE 1. Accuracy of System for Skill Search

Type	Sub Type	No.of CV's ranked	Correctly Ranked	Incorrectly Ranked	Accuracy (%)	Overall Accuracy(%)
Result when admin searches for a required skill	Python	10	9	1	90	80
	Java	10	10	10	100	
	HTML	10	10	10	100	
	Latex	10	6	4	60	
	XML	10	5	5	50	

VIII. APPLICATION

There are various applications of this CV based employability prediction system. The application are as follow,

Consultancy companies receives thousands of CV for different types of job. Among that they have to choose the perfect candidate for a particular job. This system will help the company by reduce the hard work for sorting the perfect candidate for a particular job.

It will also reduce the cost for the recruitment process and also save large amount of time.

This system can be used effectively in campus placements. In campus placements, the number of candidate are large in number so it is very difficult to sort them. This system helps to sort the candidate according to the Rank and display the selected candidate best for the position that the company has provided.

IX. CONCLUSION AND FUTURE SCOPE

In this work, the study of Machine Learning Algorithms is presented. The different techniques such as K - Nearest Neighbor, Knuth Morris Prath, etc is explained with examples. It also describes the algorithm/ steps of working with the proposed architecture. The different standard datasets is shown as input along with stages of working are defined that may be used in experiment for this system. The applications of this domain is identified and presented.

In future, a system with a even better accuracy can be developed. A larger dataset with more number of CV's can be used which would also improve the accuracy. More different criteria can be provided to the admin for a better result.

X. ACKNOWLEDGEMENT

It is a great pleasure and moment of immense satisfaction for us to express my profound gratitude to our Project Guide, Prof. Shubhangi Chavan whose constant encouragement kept us motivate enabled us to work enthusiastically. We are thankful to Prof. SatishKumar Verma, H.O.D, Information Technology Department and Dr. Sandeep Joshi, Principal, Pillai College of Engineering, New Panvel, for providing an outstanding academic environment and platform and adequate facilities. We are thankful to all our teachers who are willing to always help us whenever needed.

XI. REFERENCES

- [1] Madhavi Girase, Suchita Lad, Purna Pachpande, "Student's Employability Prediction Using Data Mining", in International Journal of Scientific Engineering Research Volume 9, Issue 4, April-2018 ISSN 2229-5518;
- [2] Neeraj Khadilkar, Deepali Joshi, "Predictive Model on Employability of Applicants and Job Hopping using Machine Learning", in International Journal of Computer Applications (0975 – 8887) Volume 171 – No.1, August 2017;
- [3] G.Vadivu, K.Sornalakshmi, "Applying Machine Learning Algorithm for Student Employability Prediction using R", in International Journal of Pharmaceutical Sciences Review and Research, Article No. 11, Pages: 38-41, April 2017;
- [4] Tripti Mishra ,Dharminder Kumar, Sangeeta Gupta, "Students' employability prediction model through data

mining", in International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 4, March 2016;

[5] Rajendra.S Choudhary, Rajul Kukreja, Nitika Jain, Shikha Jain, "Personality and Education Mining based Job Advisory System", in International Journal of Artificial Intelligence and Interactive Multimedia, Vol. 2, No 7, September 2014;

[6] Dorina Kabakchieva, "Predicting Student Performance by using Data Mining Methods for Classification ", in Cybernetics and Information Technologies • Volume 13, No 1, March 2013;

[7] Evanthia Faliagka, Kostas Ramantas, Athanasios Tsakalidis, Giannis Tzimas, "Application of Machine Learning Algorithms to an online Recruitment System", in The Seventh International Conference on Internet and Web Applications and Services, January 2012;

[8] Evanthia Faliagka, Lefteris Kozanidis, Sofia Stamou, Athanasios Tsakalidis, and Giannis Tzimas, "A personality mining system for automated applicant ranking in online recruitment systems" ;

ECG Based Heartbeat Classification

Ashwin Nair, Nikhil Elayath, Shashank Nair, Hariharan V

Department of Information Technology, PCE, Navi Mumbai, India - 410206

Abstract: The heart is a muscle that contracts in a rhythmical manner, pumping blood throughout the body. This contraction has its beginning at the atrial sine node that acts as a natural pacemaker, and propagates through the rest of the muscle. This electrical signal propagation follows a pattern. Each beat of the heart is represented on the electrocardiogram (ECG) by a wave arm. ECG can be used for detecting heart diseases. By analyzing the electrical signal of each heartbeat, it is possible to detect abnormalities. In this project, the existing methods of ECG-based automated abnormalities such as irregular heartbeat classification is surveyed. The different steps namely preprocessing, augmentation and learning algorithms are surveyed and implemented using deep convolutional neural networks to realize the use ECG signal for detecting heart diseases. The process also includes conversion of ECG signals into 2-D ECG images and designing of neural network. The dataset is identified for experiment to analyze the result using graphical user interface.

Keywords: ECG, Segmentation, Transformation, Augmentation, CNN, Activation function.

1.Introduction

Automatic ECG signal analysis to diagnose and treat cardiac diseases is of special importance in medical science. Modeling the ECG under different circumstances is very important to understand the cardiovascular system's function and to diagnose heart disease. Arrhythmias indicate a serious threat for

patients with a history of complications, such as ventricular tachycardia (VT), ventricular fibrillation (VF), and acute myocardial infarction. Arrhythmia refers to any kind of disorder in the natural cardiac rhythm. This is important since some of ECG beats are ignored in noise filtering and feature extraction. In addition, training data can be enlarged by

augmenting the ECG images which results in higher classification accuracy. Using ECG image as an input data of the ECG arrhythmia classification also benefits in the sense of robustness.

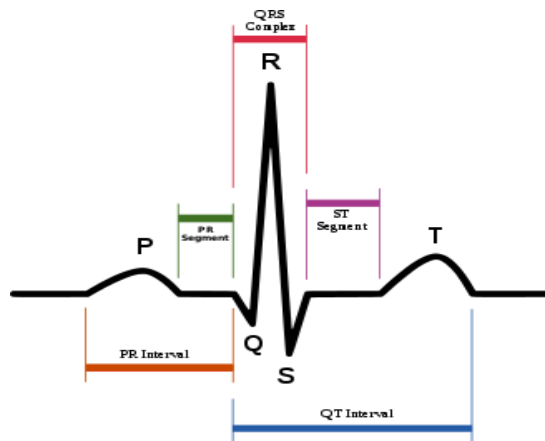


Fig. 1.1 Sample ECG signal [3]

2.Literature Survey

A. Novruz Allahverdi et al. [1] proposed a transfer learning approach for Arrhythmia Detection and Classification in Cross ECG Databases. This approach relies on a deep convolutional neural network (CNN) pretrained on an auxiliary domain (called ImageNet) with very large labelled images coupled with an additional network composed of fully connected layers. As the pretrained CNN

accepts only RGB images as the input, we apply continuous wavelet transform (CWT) to the ECG signals under analysis to generate an over-complete time–frequency representation. Then, we feed the resulting image-like representations as inputs into the pretrained CNN to generate the CNN features. Next, we train the additional fully connected network on the ECG labeled data represented by the CNN features in a supervised way by minimizing cross-entropy error with dropout regularization.

B. Eduardo Jose et al. [4] they surveyed the current state-of-the-art methods of ECG-based automated abnormalities heartbeat classification by presenting the ECG signal preprocessing, the heartbeat segmentation techniques, the feature description methods and the learning algorithms used. In addition, we describe some of the databases used for evaluation of methods indicated by a well-known standard developed by the Association for the Advancement of Medical Instrumentation (AAMI) and described in ANSI/AAMI EC57:1998/(R)2008 (ANSI/AAMI, 2008).

C. Quazi Abidur Rahman et al [3] proposed that the classifier's underlying task is to recognize individual heartbeats segmented from 12-lead ECG signals as HCM beats, where heartbeats from non-HCM cardiovascular patients are used as controls. This paper presents a cardiovascular-patient classifier we developed to identify HCM patients using standard 10-second, 12-lead ECG signals. Hypertrophic cardiomyopathy (HCM) is a cardiovascular disease where the heart muscle is partially thickened and blood flow is (potentially fatally) obstructed. Patients are classified as having HCM if the majority of their recorded heartbeats are recognized as characteristic of HCM.

D. Taotao ZHU et al. [2] proposed the study an accurate method for patient-specific ECG beat classification. adopts morphological features and timing information. As to the morphological features of heartbeat, an attention-based two-level 1-D CNN is incorporated in the proposed method to extract different grained features automatically by focusing on various parts of a heartbeat. As to the timing information, the

difference between previous and post RR intervals is computed as a dynamic feature. Both the extracted morphological features and the interval difference are used by multi-layer perceptron (MLP) for classifying ECG signals. In addition, to reduce memory storage of ECG data and denoise to some extent, an adaptive heartbeat normalization technique is adopted which includes amplitude unification, resolution modification, and signal difference.

E. An algorithm based on wavelet transforms (WT's) has been developed by Li Zheng et al. [5] for detecting ECG characteristic points. With the multiscale feature of WT's, the QRS complex can be distinguished from high P or T waves, noise, baseline drift, and artifacts. The relation between the characteristic points of ECG signal and those of modulus maximum pairs of its WT's is illustrated. By using this method, the detection rate of QRS complexes is above 99.8% for the MIT/BIH database and the P and T waves can also be detected, even with serious baseline drift and noise.

F. In the paper proposed by Martínez, J.P et al. [6], the development and evaluation

of a robust single-lead electrocardiogram (ECG) delineation system based on the wavelet transform (WT) is done. In the first step, QRS complexes are detected. Then, each QRS is delineated by detecting and identifying the peaks of the individual waves, as well as the complex onset and end. Finally, the determination of P and T wave peaks, onsets and ends is performed. We evaluated the algorithm on several manually annotated databases, such as MIT-BIH Arrhythmia, QT, European ST-T and CSE databases, developed for validation purposes.

G. In the paper, Arrhythmia beat classification using pruned fuzzy k-nearest neighbor classifier [7], Pruned Fuzzy K-nearest neighbor (PFKNN) classifier is proposed to classify different types of Arrhythmia beats present in the MIT-BIH Arrhythmia database. Fuzzy KNN (FKNN) can be implemented very easily but large number of training examples used for classification which can be very time consuming and requires large storage space. Hence, proposal of a time efficient pruning algorithm especially suitable for FKNN which can maintain good classification accuracy

with appropriate retained ratio of training data is done.

H. Emerging work with rectified linear (ReL) hidden units in the paper Rectifier Nonlinearities Improve Neural Network Acoustic Models [8] demonstrates additional gains in final system performance relative to more commonly used sigmoidal nonlinearities. In this work, exploration of the use of deep rectifier networks as acoustic models for the 300 hour Switchboard conversational speech recognition task is performed. Using simple training procedures without pretraining, networks with rectifier nonlinearities produce 2% absolute reductions in word error rates over their sigmoidal counterparts. Analysis of hidden layer representations to quantify differences in how ReL units encode inputs as compared to sigmoidal units.

I. In this paper, Thomas Unterthiner et al. [9] introduces exponential linear unit (ELU) speeds up learning in deep neural networks and leads to higher classification accuracies. Like rectified linear units (ReLU), leaky ReLUs (LReLU) and parametrized ReLUs (PReLU), ELUs alleviate the vanishing

gradient problem via the identity for positive values. However, ELUs have improved learning characteristics compared to the units with other activation functions. In contrast to ReLUs, ELUs have negative values which allows them to push mean unit activations closer to zero like batch normalization but with lower computational complexity.

J. Introduction of Adam [10], an algorithm for first-order gradient-based optimization of stochastic objective functions, based on adaptive estimates of lower-order moments. The method is

straightforward to implement, is computationally efficient, has little memory requirements, is invariant to diagonal rescaling of the gradients, and is well suited for problems that are large in terms of data and/or parameters. The method is also appropriate for non-stationary objectives and problems with very noisy and/or sparse gradients. The hyper-parameters have intuitive interpretations and typically require little tuning. Some connections to related algorithms, on which Adam was inspired, are discussed.

Table 2.1 Comparative study of techniques

Authors	Features	Future Scope
Novruz Allahverdi et al September 2016 .[1]	ECG waveforms were extracted from long-term ECGs using the moving window analysis technique.Has an accuracy rate ranging from 95% to 99%	Demanding process for clinicians and also for computer-aided systems.
Quazi Abidur Rahman et al. 2015 [3]	Uses adaptive filters based on neural networking thus significantly reducing the noise in the signals. .	Does not use normalised RR thus affecting the classification results and giving an accuracy rate of only 79%

David Menotti Nov. 2011 et al. [2]	Based on digital filters for the attenuation of the noise and removal of the fluctuating baseline.	Main goal of this work is to evaluate the benefits of using an ensemble of SVMs for the heartbeat classification problem.
Xiaoyan et al. 2007 [4]	An attention based two-level 1-dcnn is incorporated in the proposed method to extract different grained features automatically.	Feature extraction is done by focusing on various parts of heartbeat
Li Zheng et al. January 1995 [5]	By using this method, the detection rate of QRS complexes is above 99.8% for the MIT/BIH database and the P and T waves can also be detected, even with serious baseline drift and noise.	The relation between the characteristic points of ECG signal and those of modulus maximum pairs of its WT's is illustrated.
Martínez, J.P April 2004 [6]	The mean error obtained with the WT approach was found not to exceed one sampling interval, while the standard deviations were around the accepted tolerances between expert physicians.	Outperformance of the results of other well known algorithms, especially in determining the end of T wave.
Arif M et al. (2009) [7]	Proposal of a time efficient pruning algorithm especially suitable for FKNN which can maintain good classification accuracy with	By using the pruning algorithm with Fuzzy KNN, achievement of beat classification accuracy of 97% and geometric mean of sensitivity is 94.5% with only

	appropriate retained ratio of training data.	19% of the total training examples.
Andrew L. Maas et al. [8]	Using simple training procedures without pretraining, networks with rectifier nonlinearities produce 2% absolute reductions in word error rates over their sigmoidal counterparts.	Analysis of hidden layer representations to quantify differences in how ReL units encode inputs as compared to sigmoidal units.
Thomas Unterthiner et al. [9]	ELUs code the degree of presence of particular phenomena in the input, while they do not quantitatively model the degree of their absence.	In experiments, ELUs lead not only to faster learning, but also to significantly better generalization performance than ReLUs and LReLU on networks with more than 5 layers.
Diederik P. Kingma et al. [10]	The method is straightforward to implement, is computationally efficient, has little memory requirements, is invariant to diagonal rescaling of the gradients, and is well suited for problems that are large in terms of data and/or parameters.	Empirical results demonstrate that Adam works well in practice and compares favorably to other stochastic optimization methods.

3. Proposed methodology

ECG based heartbeat classification requires various steps that have to be implemented on the input values and the parameters.

3.1 Proposed system architecture

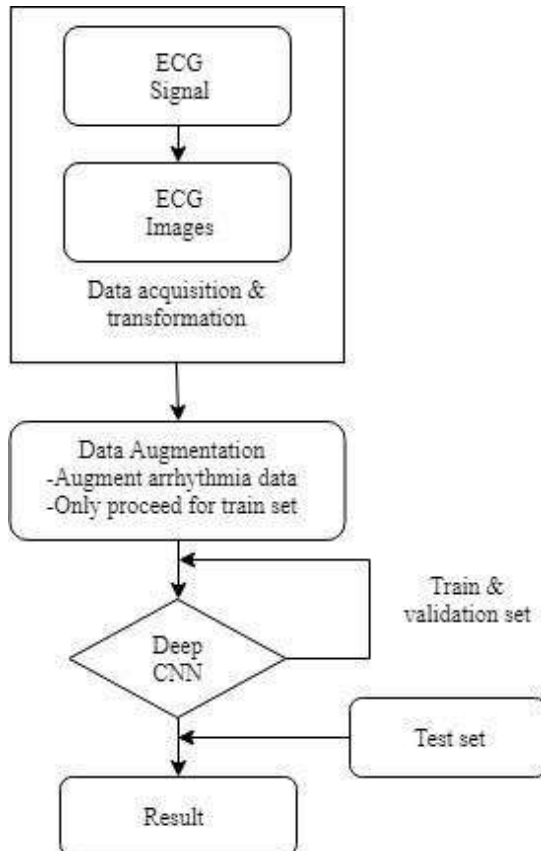


Fig. 3.1 Architecture of our project

3.2 Methodology

The following procedures are important optimization techniques that we considered while constructing the proposed CNN model.

Data augmentation

Data augmentation is one of the key

benefits of using images as input data. The majority of previous ECG arrhythmia works could not manually add an augmented data into training set since the distortion of single ECG signal value may downgrade the performance in the test set.

Kernel Initialization

The main pitfall of gradient descent based learning is that the model may diverge or fell into a local minimum point. Therefore, intelligent weight initialization is required to achieve convergence. In CNN, these weights are represented as kernels (or filters) and a group of kernels forms a single convolution layer. The proposed CNN model uses the Xavier initialization. This initializer balances the scale of the gradients roughly the same in all kernels.

Activation function

The role of activation function is to define the output value of kernel weights in the model. In modern CNN models, nonlinear activation is widely used, including rectified linear units (ReLU), leakage rectified linear units (LReLU)[8], and exponential linear units (ELU)[9]. Although ReLU is the most widely used

activation function in CNN, LReLU and ELU provide a small negative value because ReLU translates whole negative values to zero, which results in some nodes no longer participate in learning.

Regularization

Regularization also called normalization, is a method to reduce the overfitting in the training phase. Typical normalization methods are L1 and L2 normalization, however, it is common to apply dropout and batch normalization in recent deep CNN models. In deep learning, when a layer is deepened, a small parameter change in the previous layer can have a large influence on the input distribution of the later layer. This phenomenon is referred to as internal covariate shift.

Cost and optimizer function

The cost function is a measure of how well the neural network is trained and represents the difference between the given training sample and the expected output. The cost function is minimized by using optimizer function. There are various types of cost functions, but deep learning typically uses a cross-entropy function.

$$C = - \frac{1}{n} \sum [y \ln a + (1 - y) \ln(1 - a)] \quad (6)$$

Where n is the number of training data (or the batch size), y is an expected value, and a is an actual value from the output layer. To minimize the cost function, a gradient descent-based optimizer function with a learning rate is used.

Validation set

The validation set is used to determine whether a model has reached sufficient accuracy with given training set. Without the validation procedure, the model is likely to fall overfitting. Generally, validation criterion for the CNN is the loss value. However, according to observation, early stopping the model based on loss value could not achieve higher sensitivity in seven arrhythmia classification.

4. Comparative study of papers

4.1 Methodologies incorporated

The papers surveyed incorporates different methodologies depending upon their approach towards their goal.

Following are the methodologies that have been implemented on the papers that have been surveyed.

	Noise filtering	Augmentation
[1]	✓	✓
[2]		✓
[3]	✓	
[4]		
[5]		✓
[6]	✓	
[7]	✓	✓
[8]		
[9]	✓	
[10]		✓

	Efficiency	Adaptive filters
[1]	✓	
[2]		✓
[3]	✓	
[4]	✓	
[5]		
[6]	✓	
[7]		✓
[8]	✓	
[9]		
[10]	✓	

Acknowledgement:

We are profoundly grateful to Dr. Satishkumar L. Varma for his guidance and continuous encouragement throughout the completion of the synopsis.

We would like to further thank Dr. Sharvari Govilkar, Head of Department of Information technology and Dr. Madhumita Chatterjee, Head of Department of Computer Engineering and the faculty for providing us to pursue this project and also in helping us to guide and decide between plethora of options the college had to offer.

We would like to thank Dr. Sandeep M. Joshi for providing the required resources and guidance for the project. Without his support this project would not have been possible and we are grateful for his encouragement.

References:

[1] Yakup Kutlu, Gokhan Altan, Novruz Allahverdi, Arrhythmia classification using waveform ECG signals, 3rd International Conference on Advanced Technology & Sciences (ICAT'16), pp:240-245, At Konya, Turkey. September 2016

[2] Yande XIANG, Jiahui LUO, Taotao ZHU, Sheng WANG, s, Xiaoyan XIANG, and Jianyi MENG, ECG-Based Heartbeat Classification Using Two-Level Convolutional Neural Network and RR Interval Difference, 2018 Volume E101.D Issue 4 Pages 1189-1198.

[3] Quazi Abidur Rahman, Larisa G. Tereshchenko, Matthew Kongkatong, Theodore Abraham, M. Roselle Abraham, M. Roselle Abraham, Utilizing ECG-Based Heartbeat Classification for Hypertrophic Cardiomyopathy Identification, 2015 Apr 24.

[4] Eduardo José da S. Luza William, Robson Schwartz, Guillermo, Cámara-Chávez, David Menotti, ECG-based heartbeat classification for arrhythmia detection: A survey, 17 December 2015

[5] C. Li, C. Zheng, C. Tai, Biomedical Engineering Institute of Xi'an Jiaotong University, Xi'an, Shaanxi, 710049, China, Detection of ECG Characteristic Points Using Wavelet Transforms,

[6] Martínez, J.P., Almeida, R., Olmos, S., Rocha, A.P., Laguna, P., A Wavelet-Based ECG Delineator Evaluation on Standard Databases. 2004 Apr, pg no 570-81.

[7] Arif M, Akram MU, Afsar FA, Arrhythmia beat classification using pruned fuzzy k-nearest neighbor classifier. International Conference of Soft Computing and Pattern Recognition pp 37-42. (2009)

[8] Andrew L. Maas, Awni Y. Hannun, Andrew Y. Ng, Rectifier Nonlinearities Improve Neural Network Acoustic Models. in ICML Workshop on Deep Learning for Audio, Speech and Language Processing (2013)

[9] Djork-Arné Clevert, Thomas Unterthiner, Sepp Hochreiter, Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). 23 Nov 2015.

[10] Diederik P. Kingma, Jimmy Ba, Adam: A Method for Stochastic Optimization, 22 Dec 2014

Employee tracking and attendance management system using RFID.

Priya Rajput, Sofiya Rao, Laxmi Tanavarappu, Manali Thombare and Prof. Krishnendu Nair

Department of Information Technology, PCE, Navi Mumbai, India - 410206

Abstract— RFID Based Attendance System is a system developed for daily employee attendance in an organization. Employee's proper attendance management is till date a critical issue in many organization. The ability of system to uniquely identify each person based on their RFID tag (ID card) make the process of taking the attendance easier, faster and secure as compared to conventional method. To prevent fake entries by employees, we are developing a system which will be capable to track & accordingly generate the salary of employee. The functionalities of the system include senior authority tracking their associated employee, the salary message & email generation at the end of every month sent to the employee, calculation of working hours, warning message sent to employee if working hour's policy is violated. The employee only need to place their ID card on the reader and their attendance will be taken immediately. With real time clock capability of the system, attendance taken will be more accurate. The system can be connected to the computer through RS232 or Universal Serial Bus (USB) port and store the attendance taken inside database.

Keywords—

1. RFID reader
2. RFID tags
3. Attendance management
4. Salary generation
5. Employee tracking.

1. Introduction

The employee attendance tracking is the vital part of the organization. The most common means of tracking the employee attendance in the organization is by enforcing the employees to manually sign the attendance daily. There are tremendous disadvantages of using such system. For example, in large organizations it is difficult to track the attendance manually, the chances of fraud entries are more and even there are the chances of loss of the hard copy of attendance sheet.

To overcome the drawbacks of traditional manual attendance tracking system we intend to develop a system that records the employee attendance using RFID (Radio Frequency Identification). Its ability to uniquely identify each person based on their RFID tag (type of ID card) make the process of taking the attendance easier, faster and secure as compared to conventional method. The employee only need to place their ID card on the reader and their attendance will be taken immediately. The purpose of this project is to develop an attendance management system to maintain a system that helps in organization of salary, regularity and punctuality of the employee. The use of RFID technology enables the organization management to avoid attendance forms from damages such as tear, lost, and misplaced. With real time clock capability of the system, attendance taken will be more accurate and the arrival time and departure time will be automatically messaged to the employee. The system can be connected to

the computer through RS232 or Universal Serial Bus (USB) port and store the attendance taken inside database.

2. Literature Survey

A. An automatic attendance monitoring system using RFID and IOT using cloud.

In this project, RFID tags embedded in student ID cards which possesses a unique id number tagged to that student. Using readers these id numbers are scanned and student attendance is recorded. A Wi-Fi adapter is used to transfer data from reader to cloud. Here cloud is used to store data because maintaining a computer for the server needs to be kept active always. Which increases the cost, so they use cloud for their database to decrease the maintenance cost. A real time video taken at the beginning of the lecture is used to extract a single frame from a continuous set of frames and compared with existing college database which already has student pictures stored. Image comparison is done using naive similarity algorithm.

B. RFID-Based Attendance Management System.

This paper gives an implementation of event attendance tracking using RFID technology. Here the RFID reader is installed at every room of a professional event and a server application on a laptop to collect and process information. The server tasks include collection and processing of information and displaying all data via GUI in real time and store data into Ms Excel Database for analysis by even organizers or other managers. A wireless router bridges all communications between RFID reader and server. This paper gives more information about the different hardware types that can be used to implement the system with less cost.

C. RFID based Student Monitoring and Attendance Tracking System.

In this paper, implementation is divided into 4 units which are main system, receiver, transmitter and GSM module. The main system includes an interface between the computer and the receiver circuit. It also latches the data. Receiver are placed in three different zones covering area of 15 feet suppose if the area is more than 15 feet's then 2 receivers can be placed in same zones. These receivers take the location of the I-cards. Transmitters are the I-cards which transmit their location. These are active RFID cards which tend to enable multiple tags to be within range of a reader by use of "handshaking" between tags and readers. These active tags are much faster than in passive tags. GSM module sends messages of the data entered. This module works on the AT-commands and these commands can be programmed in visual basics. GSM/GPRS Modules are alike to modems, but there's one difference: A GSM/GPRS Modem is outer equipment, whereas the GSM/GPRS Module is a module that can be integrated within apparatus. It is an embedded piece of hardware.

D. Smart Attendance System by suing RFID

This paper mainly emphasizes on the GUI and optimality of the tracking system using the basic programming methodologies. There is a MySQL Database used to provide data support and backend is developed using PHP. All the functionalities can be achieved using the integrated RFID Database handling system. The SAS fetched the user data such i.e, the reader ID from RFID database and match with the tag run a set of queries and executes the attendance tracking process. Problems associated with this is that no proper mechanism to implement the RFID tag-reader module is given. The emphasis is more on the GUI and less on the actual working of the system

3. Proposed Work

In our proposed system, we are taking into consideration the need for improvement of an employee attendance system by increasing its scope of functionality as an outcome of embedment of additional modules into the existing architecture. The enhancements are as follows:

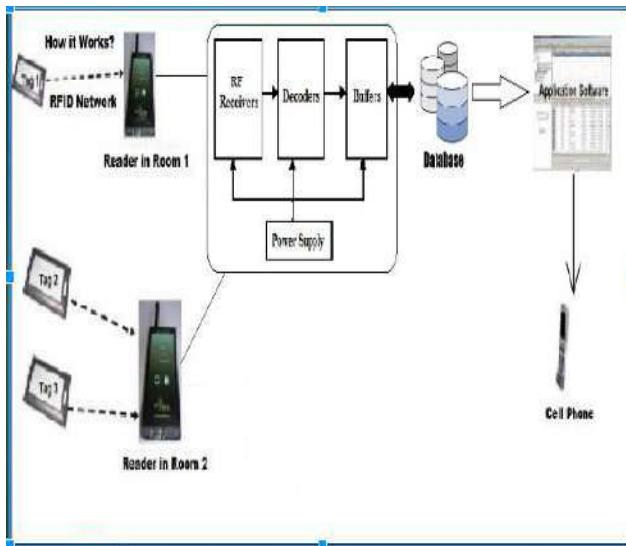


Fig 1:-Proposed system architecture

In the above block diagram it shows that there are 3 different tags and each tag is handed to the person as the ID card. Receivers are placed in three different zones suppose office area, canteen. Each receiver will cover the area of 15 feet's suppose if the area is more than 15 feet's then 2 receivers can be placed in same zones. When the person carrying the tag comes in the ranges of 15 feet's the receivers will sent the data to the decoder of the mother circuit. Then this data is given to buffer after going through the buffer the output of the buffer is given to the parallel port. Parallel port is connected to the PC. PC is used for storing the record of all the tags also it is used to monitor the on screen movement of the tag from one zone to another also to calculate the time he/she was present in particular zone. Also the records of any date can be known from the database.

1.Main System

The RFID reader scan the RF signal from the RFID tags and transmit the resulting tag signal at fixed intervals. The antennas receive and process the response. The Reader transmits the data from the antenna to a host computer. The host computer assembles the data and resolves them into positional estimates. Data are archived in a data warehouse, such as an access database and accordingly message is sent and using the same data the salary is generated at the end of every month.

2. Receivers

The signals transmitted at fixed interval are collected by the receiver and sores into the data warehouse for further use. The data stored in the data ware house are used to send messages to each employee about the in and out time and accordingly calculate the salary at the end of the month.

3. Tags (Id-cards)

Active Tag systems require battery-powered tags. The battery admits a longer detection range of between 3 and 100 meters. These systems are able to locate the tags with higher accuracy than passive RFID systems and typically operate in the 400, 900, or 2440-megahertz bands. Active tags tend to enable multiple tags to be within range of a Reader by the use of "handshaking" between the tags and Reader, so that each tag transmits its signal in turn. Communication between tag and Reader in active systems is also typical faster than with passive tags.

3.1 System Architecture

The overall architecture of the system is illustrated in Figure, where the three main components are shown. Each of these components will be described in the following sub-sections.

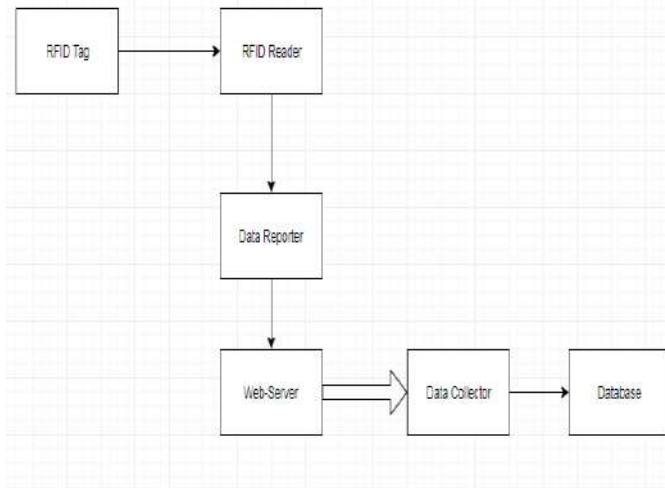


Fig 2:-Existing System Architecture

A. RFID Reader and Tag

RFID reader is the device capable of reading and retrieving information stored inside the RFID tags. There are two types of RFID reader, which are the active and passive RFID readers. Active RFID reader can detect an active RFID tag while passive RFID reader can only detect passive RFID tag at a few centimeters away from the reader. The RFID reader being used in the system is a low cost reader for reading passive RFID tags. It operates at 0~400C temperatures, 20~80% of humidity, 125 kHz frequency and 12V power supply. The effective detection range of the reader is around 5-8cm. Each RFID tag has a unique serial number or ID. There are three types of RFID tags which are active, semi-passive and passive. The main difference between these RFID tags is that active and semi-passive RFID tags require internal battery while passive RFID tags do not use any internal battery. Adapted to our scope of work, the employee cards being used to identify each individual employee are the RFID cards that consist of passive RFID tag, which do not require internal battery. When such cards are passed through the field generated by a compatible Reader, they transmit information back to the Reader[12]

B. Data Reporter

Data Reporter is a component that fetches all logging data from the RFID reader such as the captured employee ID, time and date for some interval. The collected data are then passes to the online server, which will record the data into the database. This component should always be kept up and running and needs to be automatically restarted each time the operating system reboots[12].

C. Web Server

The web server here refers to either hardware (computer) or software (application) that helps to deliver content publicly accessible through the Internet. It provides the web site functionality by accepting requests from the user's browser and responds by sending back HTML documents (Web pages) and files. To enable the system dynamic functionalities, the web server hosts the data collector component, a database and the graphical user interface (GUI) pages enabling online interaction with the system users.

D. Data Collector

The role of the online data collector is to continually listen to incoming data sent by the Data Reporter component. The received log data will then be inserted to the database for recording purpose.[12]

E. Database

A database is defined as an organized collection of data and tailored to our system, our database is employed to mainly store the data captured by the RFID reader. Secondly the database is also used to store data gathered from the online web-interface, such as class schedule and employee personal information. In offering more features to the users, our online- system can manipulate the recorded employee attendance record by querying the database for complex data retrieval [14]. This includes automated

operation, such as summarizing an individual student attendance by calculating the attendance

3 Requirement Analysis

The implementation detail is given in this section.

3.1 Software

Processor	2 GHz Intel
HDD	180 GB
RAM	8 GB
RFID Reader & Writer	

It is our privilege to express our sincerest regards to our supervisor Prof. Krishnendu Nair for the valuable inputs, able guidance, encouragement, whole-hearted cooperation and constructive criticism throughout the duration of this work. We deeply express our sincere thanks our Head of the Department Dr. Sharvari Govilkar and our Principal Dr. Sandeep M. Joshi for encouraging and allowing us to presenting this work.

REFERENCES

- [1.] An automatic attendance monitoring system using RFID and IOT using cloud. Author:-Tarun Sharma and S.L Aarchy Date:- 4 May 207
- [2] Automation of attendance system using RFID, Biometric, GSM Modem with .NET framework. Author: Aamir Nizam Ansari, Arundhati Navada, Sanchita Agarwal, Siddharth Patil, Balwant A. Date: 2011

3.2 Hardware

Operating System	Windows XP Professional With Service pack 2
Programing Language	JDK 1.8
Database	Oracle 9

- [3.] Building a smart university using RFID technology. Author: Aqeel-er-Rehman, Abu Zafar, Zubair.A.Shaikh Date: 2008
- [4]. Smart Attendance System by using RFID. Author: M. K. Yeop, M. Z. A. Abdul Aziz, M. S. R. Mohd Shah, M. F. Abd Kadir Date: 22 August 2008
- [5.] <https://www.youtube.com/watch?v=Ukfpq71BoMo>

ACKNOWLEDGMENT

IOT BASED ENERGY MONITORING SYSTEM

Sameer Mhatre, Vishal Patil, Jaiprakash Khichi, Chandan Chodhary, and Guide Prof. K.S.Charumathi

Department of Information Technology, PCE, Navi Mumbai, India - 410206

Abstract—Nowadays Energy Monitoring & Preservation hold prime importance in day by day life. In market there are many electronic energy monitoring system are available. Most of them are monitor the power consumed in a residence household. Many a time, the consumer are not satisfied with electricity bill as the device does not show the power consumed at device level. The IOT based devices created a revolution in electronic & IT. The proposed system is designed such that can measure the power of consumption by an individual electrical appliance. The proposed system uses an energy meter using Arduino Microcontroller. The main purpose of energy meter is to measure the power consumption at the device level, upload it to the server. Energy monitoring system measure the power consumed by various electrical devices & display it to energy monitoring website. The advantage of energy monitoring system is that a user understand the power consumed by each devices in a particular premises in order to control the power consumption & used by each devices as well as to save energy by controlling them.

1. Introduction

Currently, the energy is one of the most important need in this days. The idea of energy efficient device has come from various areas such as air conditioning, lighting etc. The energy bill is generated on monthly basis that the user can analysing the power consumption detail in

A. IOT Based Smart Energy Monitoring. In the present system, energy load consumption is

every month. Energy meter is installed in the residential house, building that shows the consumption of energy in household. In this paper we are implementing an energy monitoring system that display the power consumption of single or multiple devices. So the user can detect any error in the bill. The main objective of the project is to monitor power consumption of each and every electrical device to reduce the monthly electricity bill. The data of power consumption of devices will help to lower the usage of the devices that are not in use and hence increasing the life of the device. This project allows people to monitor their household power consumption on daily basis. By regular monitoring, the energy would be saved to some extent. People will be able to monitor energy usage of the devices that consume excess energy, so by reducing the usage of those devices that are not in use and still consuming energy will help to lower the monthly energy bill. A relay would be used to cut off the power of the devices that reach maximum power units that has been set in the energy monitoring system.

2. Literature Survey

We have surveyed a variety of papers on energy monitoring system. These papers discuss the basic architecture in which readings are taken from sensors and are display on LCD Board. the devices are characterized by easy access to the information and combination of smart meter and data communication capability allows local and remote access.

accessed using Wi-Fi and it will help consumers to avoid unwanted use of electricity. IoT system

where a user can monitor energy consumption and pay the bill Online can be made. Also, a system where a user can receive SMS, when he/she crosses threshold of electricity usage slab can be equipped. We can make a system which can send SMS to the concerned meter reading man of that area when theft is detected at consumer end. Also using cloud analytics we can predict future energy consumptions.

B. IoT Based Smart Energy Management System. The present IoT based power management systems using image processing is very costly, hence the paper proposes a cheap method to regulate the wastage of power by giving penalty to the individual or organization by the power distributors. The system comprises of thermal sensing and associated hardware, this system requires less installation cost and maintenance is cheap. This method allows the government to control the entire power wastage by remotely and it helps in power saving.

C. IOT Based Smart Energy Management System. So with all these work reported, we here have developed an better IoT system for Energy Management which takes the Humidity, Temperature and light intensity into consideration and accordingly interfaced with Arduino Microcontrollers for controlling the usage of appliance like speed of fan, light intensity rather than just switch on or off. Also the prototype system computes the current drawn from each appliance based on appliance usage and send to Raspberry Pi3 where total power consumed of appliances computed against time. This information is computed all through the day and same uploaded in cloud server too.

D. IOT Based online Energy Monitoring System. The goal of this project is to

visualize and monitor the power consumption online on a smart phone using mobile application by integrating smart plugs, sensors, IOT devices and GATEWAY which enables the communication between the various smart plugs and the web server hosting.

3.1 System Architecture

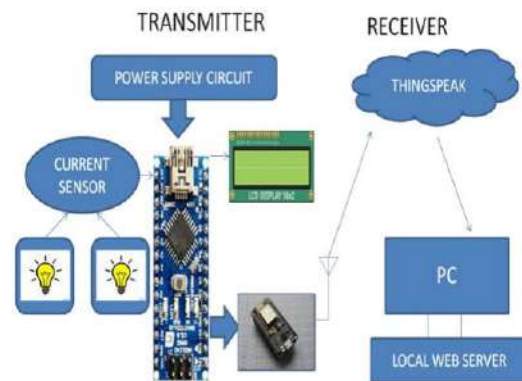
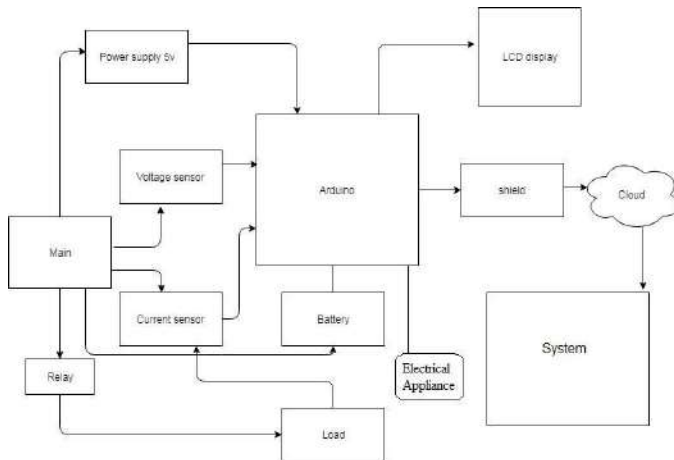


Fig:1 Existing system architecture for IOT based energy monitoring system.

In recent years, the demand of building automation system increases especially in offices and households. Generally, it is because automation helps reducing consumption of electricity, decreases the wastage, uses less manpower, and helps in energy saving. Automation system that is implemented at home is known as home automation. The home automation term is referred to the automation system that can integrate household activities which include sensors to read input condition and centralized the control of electrical appliances. Nowadays, many researchers have innovated technology with home automation. The purpose of this system is to monitor the consumption of energy by

particular appliances, like how many memory consumed by particular mobile application.

3.2 Proposed system architecture



Arduino board: Arduino is a microcontroller board and it is based on the AT mega 328P. It consists of 14 digital I/O pins and 6 analog input pins and a crystal oscillator of 16 MHz frequency, a power supply jack and a USB port to dump the code, ICSP header and a reset button. It can be powered with the power jack at the start and later can be powered with AC to DC adapter or with a battery.

wifi module: The ESP 8266 Wi-Fi module is a low cost component with which manufacturers are making wirelessly networkable microcontroller module. ESP 8266 WiFi module is a system-on-a-chip with capabilities for 2.4GHz range. It employs a 32 bit RISC CPU running at 80 MHz. It is based on the TCP/IP (Transfer control protocol) [3]. It is the most important component in the system as it

performs the IOT operation. It has 64 kb boot ROM, 64 kb instruction RAM, 96 kb data RAM. Wi-Fi unit performs IOT operation by sending energy meter data to webpage which can be accessed through IP address. The TX, RX pins are connected to the 7 and 8 pins of the Arduino microcontroller.

Current sensor: The Allegro ACS712 provides economical and precise solutions for AC or DC current sensing in industrial, commercial, and communications systems. The device package allows for easy implementation by the customer. The device is not intended for automotive applications. The device consists of a precise, low-offset, linear Hall circuit with a copper conduction path located near the surface of the die. Applied current flowing through this copper conduction path generates a magnetic field which the Hall IC converts into a proportional voltage. Device accuracy is optimized through the close proximity of the magnetic signal to the Hall transducer. A precise, proportional voltage is provided by the low-offset, chopper-stabilized BiCMOS Hall IC, which is programmed for accuracy after packaging. The output of the device has a positive slope ($>V_{IOUT}(Q)$) when an increasing current flows through the primary copper conduction path (from pins 1 and 2, to pins 3 and 4), which is the path used for current sampling. The internal resistance of this conductive path is 1.2 m Ω typical, providing low power loss.

Voltage Sensors: Arduino Voltage Sensor 0-25V. The reason I'm making this is because I couldn't find any really helpful

information on how to fix the code for my voltage sensor. Arduinos have built in voltage sensors. Unfortunately, they only support voltages of 0-5V. This module allows you to measure voltages of 0-25V by presenting a lower voltage to the arduino for measuring. After you have this value you simply feed it through some math and you get your actual voltage. Don't ask me how this math works. I don't know. If you do know however, please share. I'm really just editing the example code from the seller so that it will display decimal values instead far less useful int values. To start you need to wire it up. It's extremely easy as it only needs 3 wires. Plug + into 5V, - ground and S into an analogue pin. I have removed all but the relevant pins in a pinout of the arduino nano. If you're using another model then you'll have to figure them out on your own. Any analogue pin will do. As far as I am aware at least. Once you have done this you're ready to move on to the software.

Thingspeak graphics user interface:The Internet of Things provides access to a broad range of embedded devices and web services. ThingSpeak is an open data platform and API for the IoT that enables you to collect, store, analyze, visualize, and act on data from sensors or actuators, such as Arduino, BeagleBone Black, and other hardware. For example, with ThingSpeak you can create sensor-logging applications, location-tracking applications, and a social network of things with status updates, so that you could have your home thermostat control itself based on your current location. The primary element of ThingSpeak activity is the channel, which contains data fields, location fields, and a status field. After

ThingSpeak channel is created, you can write data to the channel, process and view the data with MATLAB® code, and react to the data with tweets and other alerts. The typical ThingSpeak workflow lets you:

1. Create a Channel and collect data
2. Analyze and visualize the data

Webpage: The proposed system can be used to display load energy usage reading in terms of Watts. Every user would be able to access the information from anywhere on the earth. Thingspeak.com is one such webpage which takes the help of the Math Works MATLAB analytics to present the device information in a more detailed analysis in both description and visualization. Thingspeak.com provides the user the ability to add any number of channels to one account and in each account information can be fed into 8 fields. An account can be assigned to one division of an area and n channels can be created to a suite of n meters in the locality. The analytics can be viewed by both the consumer and service provider.

Implementation details

Techniques:

1. Real time monitoring algorithm. Real time monitoring algorithm will take the readings of the units of energy utilised by all the devices. This reading will help us to know the power consumption of devices. Using this algorithm we can put threshold limit to the circuits which will automatically reduce the usage of the device after the threshold point is reached. Threshold point will notify the user that monthly power

consumption has reached its limits and needs to take control of the devices that are not in use.

2. Session algorithm

Session algorithm is an algorithm where a task is done in a particular time period. In this project, session algorithm would be used by the user when he needs to check the unit of power that is consumed by the devices. Each and every user will have a login id. The user can see the readings of the unit at anypoint of time after login into the website.

3.2.1 Hardware and Software Specifications

The experiment setup is carried out on a computer device for programming the arduino and on mobile device to demonstrate the mobile application, both the devices have the following hardware and software applications as shown in the below Table 3.1 and Table 3.2 respectively

Table 3.1 Hardware details

Processor	Pentium(R) Dual core CPU,2.10 Ghz
HDD	320 GB
RAM	Minimum 500 MB

Table 3.2 Software details

Operating System	Windows, Linux, IOS
------------------	---------------------

Programming Language	HTML5, CSS3, java script, Arduino
Database	Sql Server 2013

ACKNOWLEDGMENT

It is a great pleasure and moment of immense satisfaction for us to express my profound gratitude to our dissertation Project Guide, **Prof. K.S.Charumathi** whose constant encouragement enabled us to work enthusiastically. Her perpetual motivation, patience and excellent expertise in discussion during progress of the dissertation work have benefited us to an extent, which is beyond expression. We are highly indebted to his invaluable guidance and ever-ready support in the successful completion of this dissertation in time. Working under his guidance has been a fruitful and unforgettable experience. Despite of his busy schedule, he was always available to give us advice, support and guidance during the entire period of our project. The completion of this project would not have been possible without his encouragement, patient guidance and constant support.

We are thankful to **Dr. Sharvari Govilkar**, H.O.D, Information Technology Department and **Prof. Sushopti Gawade and Prof. Gayatri Hegde**, B.E. Project Coordinator, Pillai college of Engineering, New Panvel, for her guidance, encouragement and support during my project. We would like to mention here that she was instrumental in making available all the needed resources throughout our project. We are highly

indebted to her for her kind support. We are also thankful to **Dr. Sandeep M. Joshi**, Principal, Pillai College of Engineering, New Panvel, for his encouragement and for providing an outstanding academic environment, also for providing the adequate facilities

REFERENCES

- [1]. IOT Based Energy Monitoring, Abhiraj Prashant Hiwale, Deepak Sudam Gaikwad, Akshay Ashok Dongare, Prathmesh Chandrakant Mhatre, 2018.
- [2]. IOT Based Energy Meter Billing and Monitoring System, Sasane Nikita , Sakat Swati, Neman Shital, Prof. Vipul Ranjan Kaushik, Prof. Pallav P.K, 2017.
- [3]. IOT Based Energy Monitoring and Control Device, Madhuri G. Hiremath1 , Veeresh Pujari2 , Dr. Baswaraj Gadgaysign, 2017.
- [4] J. JeyaPadmini,K. R. Kashwan, “Effective Power Utilization and

Conservation in Smart Homes Using IoT “,2015 international conference on computation of power, energy, information and communication ,2015. [Type of medium]. Available FTP: Directory: File:

- [5] T. Guettari y, J. Boudy , BE. Benkelfat, G. Chollet,JL. Baldinger, “ Thermal signal analysis in smart home environment for detecting human presence”,1st International Conference on Advanced Technologies for Signal and Image Processing - ATSIP 2014,March 17-19, 2014, Sousse, Tunisia [Online]. Available FTP: atmnext.usc.edu Directory: pub/etext/1994 File: atmosplasma.txt.

- [6] Nihesh Rathod, Pratik Jain and Renu Subramanian,”Performance Analysis of Wireless Devices for a Campus-wide IoT Network”. The 2015 International Workshop on Wireless Network Measurements and Experimentation, 2015

IOT BASED WOMEN AND CHILDREN SAFETY SYSTEM

Shrayesh Kanade, Siddhi Morajkar, Priyanka Borse, Vrutant Mehta

Prof. Rupali Nikhare

Department of Information Technology, PCE, Navi Mumbai, India – 410206

Abstract:

Women safety in India is a big concern. Every day and every minute some women of all walks of life are getting harassed, molested, assaulted, and violated at various places all over the country. Not only women but children are also facing safety issues at schools, colleges and public places. When someone faces insecure situations, to ensure the safety, automatic detection system needs to be established.

A system that is designed merely to serve the purpose of providing security to women and children while facing social challenges. This can be done by using various sensors to precisely detect the real time situations of women and children. We are proposing a model which will help to ensure the safety of women and children all over the globe. We will be using heartbeat sensor. We are using GPS which will help to detect location of the device. GSM used in the model will be used to send alert message to guardians,

relatives. A buzzer button through which the message will be sent automatically. With this live tracking through app will be possible. We will be working on backtracking as well, to detect the last location of the person if the signal is lost. We have proposed IOT(internet of things) based device which will help to continuously monitor values of different sensors and GPS used. We are also using a camera which will capture images when the button is pressed and send those images to the receiver. The device can be disabled as and when needed. The proposed model will help monitor values of different sensors which will make it easy to reach the victim with great accuracy. The system will be very convenient and easy to use unlike the existing apps which are very obsolete.

Introduction:

The internet of things (IoT) is a network of physical devices, vehicles and other items

embedded with electronics and network connectivity which enable these objects to collect and exchange data. The internet of things, or IoT, is a system of interrelated computing devices, mechanical and digital machines, objects, animals or people that are provided with unique identifiers (UIDs) and the ability to transfer data over a network without requiring human-to-human or human-to-computer interaction. Each thing is uniquely identifiable through its embedded computing system but is able to inter operate within the existing internet infrastructure.

The IoT allows objects to be sensed or controlled remotely across existing network infrastructure, helps in creating opportunities for direct integration of physical world and computer-based systems resulting in improved efficiency, accuracy. When IoT is augmented with sensors and actuators, the technology becomes an instance of more general class of cyber physical systems. IoT provides accuracy, economic benefit in addition to reduced human intervention. Electronics miniaturization, cost of electronic components, and the trend towards wireless communications are the three main drivers for IoT. The core components of the IoT will be sensors and actuators, embedded processing, and connectivity and the cloud. Smart objects such as modern phones use sensors and actuators to interact with the real world. that distinguishes the Cyberbullying comments

from the regular ones.

Literature Survey:

1) G C Harikiran, Karthik Menasinkai, Suhas Shirol have proposed a model that consists of a Smart band integrated with Smart phone which has an added advantage so as to reduce the cost of the device and also in reduced size. The GPS and the GSM can be used of a smart phone. This also enables in reduced power use and that the watch can be installed with Bluetooth 4.0 BLE (Bluetooth Low Energy) which comes in handy for several days on a single shot of charge..Heart beat sensor gives digital output of heart beat.

2) D. G. Monisha, M. Monisha, G. Pavithra and R. Subhashini proposed a device designed with GSM, GPS, Bluetooth and RF detector . The whole device just runs with total of 12v in which 5v is enough for the ARM to process In this system, an Android Application is used to find the location and send the location to the group of people stored in the phone, SOS Message, Track your phone and additionally we used a technique of clicking the volume button..

3) Prof. R.A.Jain, Aditya Patil, Prasenjeet Nikam, Shubham More, Saurabh Totewar proposed a IOT based system for women safety by using ARM7LPC2148, Panic button,GSM, GPS,and different sensors like heartbeat,motion and temperature sensor at the transmitting end and Raspberry Pi at the

receiver end. Proposed Model is wearable model. After giving power supply to device , sensors on device will start taking readings.

4) A. Helen , M. Fathima Fathila, R. Rijwana, Kalaiselvi .V.K.G proposed a model having GPS and GSM modules. The GPS and GSM integrated with smart watches is connected via Bluetooth to the smart phone and ring the alert notification to the emergency contact and within the limited radius the police station will be found in the GPS and make a signal. Cop will be able to track the alert signal and find the location. Pulse rate sensor detection is used in this to detect the sensor when the targeted pulse rate is achieved. The temperature and the motion sensor is used to detect the condition of the user.

5) Roshni S. Sune, M. H. Nerkar system consist of the Raspberry Pi module which gets the signs from GPS system and after the Pi controller permits to send the Alert message with the location of the user to the saved predefined numbers. In order to track the pulse rate, pulse rate sensor is used to sense the heart rate and send a message. When user presses a button, automatically the auto defender system will start working. Auto defender system consists of the buzzer which will make sound so that someone in the surrounding can listen and help the user, second is shock mechanism which will stun the attacker, third is the sprinkler that will sprinkle the harmful, etching solvent like

pepper powder which will harm an attacker and still gets help.

Proposed System:

The IoT based device is much easier to access in dangerous situations than using a mobile phone. When the device is set on, it continuously sends out the longitude and latitude of the user over the internet to the server which helps in displaying the current location of the user on the map. The relative at such can view the map from the website and check whether there is any unusual route taken by the user. This feature will basically be much useful in case of children going to schools. It helps the parents to keep track of whether their child has reached home safely.

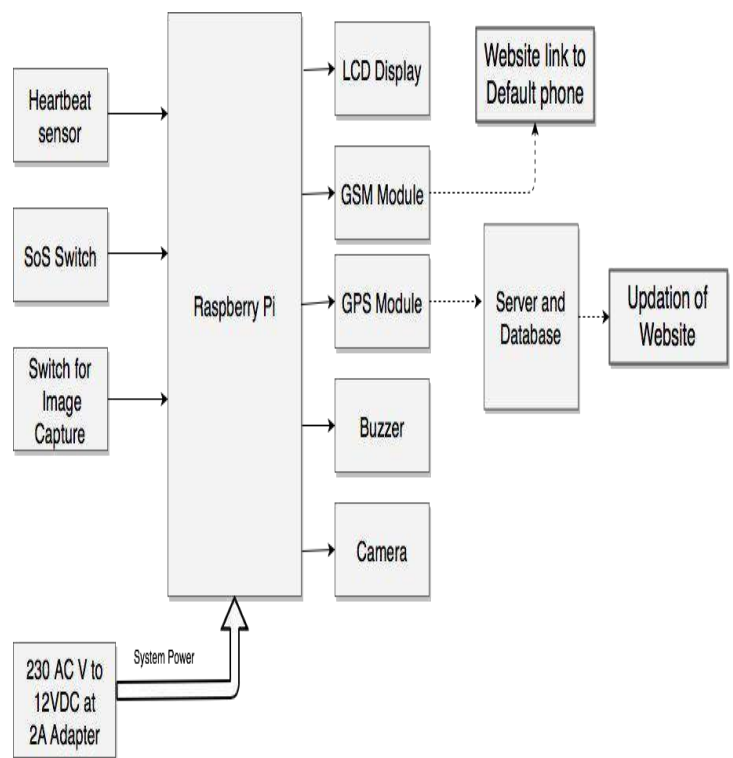


Figure 1. Block Diagram of Proposed System

Requirement Analysis

Software and Hardware Requirements

The product or tool is carried out on a system with basic minimum standards are as:

Table 1. Hardware details

Process or	Core i3 GHz or higher
HDD	180 GB min
RAM	2 GB
GSM	SIM800
GPS	Built using MT3329 chipset from MediaTek.Inc
Raspber ry Pi	B+
Camera	Logitech
Heartbe at sensor	RKI-3156
Power Supply	230 V AC to 12V DC at 2A Adapter

Table 2. Software Details

Programming Language	Python for Raspberry Pi
Front-end	HTML,CSS
Database	MySql
Backend	Php

Acknowledgement:

It is our privilege to express our sincerest regards to our supervisor Prof. Rupali Nikhare for the valuable inputs, able guidance, encouragement, whole-hearted cooperation and constructive criticism throughout the duration of this work. We deeply express our sincere thanks our Head of the Department Dr. Sharvari Govilkar and our Principal Dr. Sandeep M. Joshi for encouraging and allowing us to presenting this work.

References:

1. Vamil B. Sangoi, "Smart security solutions," International Journal of Current Engineering and Technology, Vol.4, No.5, Oct-2014.
2. B. Chougula, "Smart girls security system," International Journal of Application or Innovation in Engineering & Management, Volume 3, Issue 4, April 2014.
3. Simon L. Cotton and William G. Scanlon, "Millimeter - wave Soldier – to soldier communications for covert battlefield operation," IEEE communication Magazine, October 2009.
4. George R, Anjaly Cherian V, Antony A, et al. An intelligent security system for violence against women in public places. IJEAT; 2014 Apr 3.
5. GPS and GSM Based Self Defense System for Women Safety” Sriraniini

- R, Journal of Electrical & Electronic Systems, ISSN: 2332-0796.
6. Vijayalashmi B, Renuka S, Chennur P, Patil S (2015) Self defense system for women safety with location tracking and SMS alerting through GSM network. International Journal of Research in Engineering and Technology (IJRET) 4: 57-60.
 7. Premkumar P, Cibi Chakkaravarthi R, Keerthana M, Ravivarma R, Sharmila T (2015) One Touch Alarm System For Women's Safety Using GSM. International Journal of Science, Technology & Management 4: 1536-1539.
 8. Miriyala GP, Sunil PVVNDP, Yadlapalli RS, Pasam VRL, Kondapalli T, et al. (2016) Smart Intelligent Security System for Women. International Journal of Electronics and Communication Engineering and Technology (IJECET) 7: 41-46.
 9. Gowri Predeba B, Shyamala. N, Tamilselvi.E, Ramalakshmi.S, Selsiaulvina. "Women security system using gsm and gps" International Journal of Advanced Research Trends in Engineering And Technology (IJARTET), 2016 April.
 10. <http://www.security.honeywell.com/hsc/products/intruder-detection-systems/sensor/motion/dual-tec-commercial/790177.html>
 11. <http://chapters.comsoc.org/vancouver/BTLER3.pdf>

Literature survey on Real Time OSN Analysis To Detect Online Terrorists

Mr Mrudul Bornare^{#1}

Ms Tasneem Attarwala^{#2}

Ms Riya Jadhav^{#3}

^{#1,2} Department of Information Technology, Mumbai University

PIIT, New Panvel, India

¹manasijoshi97@gmail.com

²mrudul.bornare47@gmail.com

³tasneemattarwala97@gmail.com

⁴riyarj15it@student.mes.ac.in

Abstract-

As technology and internet advances in developing countries, OSNs have seen a major increase in their user-base. Online Social Networks(OSNs) are not only used for exchanging information but also for recruitment of people in terrorist groups which threaten the integrity of a country. The project aims to classify malicious and legitimate users based on real-time tweets extracted from twitter. These tweets are classified by using various classification algorithms like Kmeans and Naïve Bayes. Cyber criminals often exploit the social network with malicious URLs which diverts the legitimate users to a server that performs unwanted actions on the users machine. The project uses a machine classification system to distinguish between malicious and benign URLs within seconds The suspicious activity of a user is identified using semantic analysis performed on the Realtime data that is extracted from twitter

Keywords—data mining, URL detection, preprocessing, online tweets, semantic analysis, summarization.

I. INTRODUCTION

Internet and information technology are the platform where huge amount of information is available to use. This project focuses on targeting users based on their posts on the social websites. But targeting the malicious users based on their posts is tedious to deal with. N number of tweets getting uploaded every second and to retrieve useful tweets from large amount of collection from web and database may cause to miss track to user. Thus there is need to develop some approach which clearly guides the user about how to suspect the malicious user. The tweets uploaded on twitter may be in an unstructured form. Data Mining is a tool to develop effective mining algorithms to invoke particular pattern from collection of data [1], [5]. Thus to deal with such types of tweets and their result.

A person's tweet is analyzed through the above-mentioned pre processing methods which can help in finding suspects. The suspect once known can be reported on the social media sites and their information can be passed on to intelligence agencies for further investigation. In this application we are using real time online data to detect the suspects who use social media as a medium to spread terrorism. For this, we are using semantic analysis and preprocessing techniques to filter the tweets. Moreover, the information of the online suspects could be passed on to cyber crime agencies for further investigation.

In this paper, we have mentioned a literature survey on different techniques used for online terrorist detection. the relevant techniques in literature are reviewed. It describes the various techniques used in the work. It identifies the current literature on related domain problem. It also identifies the techniques that have been developed and how we work upon their limitations. It is important to know how these steps such as tokenization, morphology, stemming, filtering, semantic analysis, n gram algorithm etc. will work in a sequential format. And lastly the different techniques for the same are discussed with their methods and result that researchers found by implementing those methods.

Technological advancements have provided many means by which terrorists may misuse the Internet for illicit purposes [1]. Despite increasing international recognition of the threat posed by terrorists' use of the Internet in recent years, there is currently no universal instrument specifically addressing this problem There are some issues related to Detection of terrorist as given below.

Issue 1- The structureless data on the internet in the form of texts needs a lot of preprocessing which affects the performance of the system.

Issue 2- Sometimes the system detects a failure due to low resolution posted on social media networks..

Issue 3- The system might lose its agility due to complex network.

Issue 4- The algorithm is not up to the accuracy needed while it can still increase with certain advancements.

The project is divided into two modules. The first module consists of sentiment analysis. The second module uses sampling, classification and prediction methods. Classification and Prediction are the two forms of data analysis that can be used to predict future data trends. Data sampling is an analysis technique used to select and analyse the data further identifying patterns in the larger dataset being examined. In the first module, sentiment analysis gives the polarity of the tweets. For this initially terrorist related dataset is downloaded. SentiWordNet, a lexical resource for opinion mining, is used. It generally assigns the sentiment scores of positivity and negativity. The program consists of an array which includes terrorist related words. Based on the occurrence of such words in the real time tweets, the positive and negative values are assigned to respective tweets.

The more positive value indicates that the sentence posted by the user is in a good sense and more negative value indicates that the user is suspicious to terrorism. In the second module of the project the data containing various arabic scripts, punctuation, characters, etc. is sampled. So basically here the data is cleaned, analysed and manipulated into simpler, more neater format. Next, classification is done to accurately predict the user is suspicious or not through the tweets. This is done using a string of words. If seven words are encountered in a particular tweet then the person who tweeted is found to be suspicious. At the end prediction is done but before that the mean of all words is calculated. If two or more words from negative wordlist out of seven words in array, it shows that the person is suspicious to terrorism.

The comparative study of various techniques mentioned above is presented in this report. The performance measures like precision and recall are described in this report. The different standard datasets or variable inputs are defined that may be used in experiment for this domain systems. The applications of this domain is identified and presented. The project aims to classify malicious and legitimate users based on real-time tweets extracted from twitter.

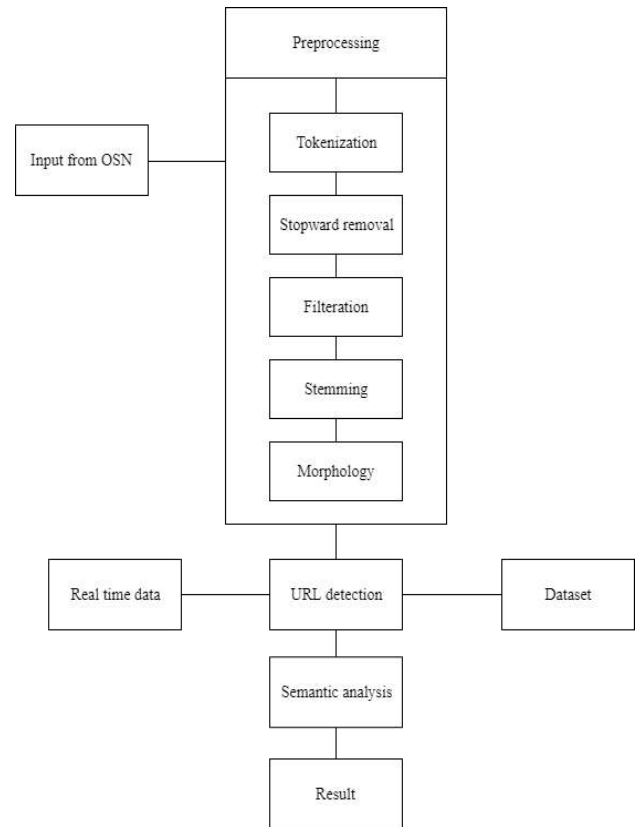


Fig.1. System Architecture For Online Terrorist Detection

In order to achieve better domain results, researchers combined both techniques to build Hybrid domain systems, which seek to inherit advantages and eliminate disadvantages.

II. TERRORIST DETECTION PROCESS ON OSN

The terrorist detection process starts by taking input from the user. Here input indicates the real time online tweets generated every second on online social network such as Twitter. After taking input from the OSN the preprocessing techniques are performed. It consists of Tokenization, Stop words removal, Filtration, Stemming and Morphology. Further the URL detection technique is used along with Semantic analysis. Following are the preprocessing techniques for text mining [3], [4] that we will discuss in detail.

- A. Tokenization
- B. Stop words Removal
- C. Filtration
- D. Stemming
- E. Morphology

A. Tokenization

In tokenization the text is split into smaller parts. These smaller parts are called as ‘tokens’. Tokenization is a crucial step in NLP. Stop words Removal

Example:

Input:

Syria Daesh claims a suicide attack in the Al-Zahra area of Homs. Three of the IS suicide bombers who attacked the military hotel in Aden are from Yemen.

Output:

“Syria” “Daesh” “claims” “a” “suicide” “attack” “in” “the” “Al” “-” “Zahra” “area” “of” “Homs” “.” “Three” “of” “the” “IS” “suicide” “bombers” “who” “attacked” “the” “military” “hotel” “in” “Aden” “are” “from” “Yemen””

B. Stopwords Removal using NLTK

When indexing entries for searching, the search engine is programmed to ignore words like ‘the’, ‘an’, ‘in’, ‘a’ which are useless when it comes to retrieving the result of search query. Such words are called as Stop words. This technique is widely used in NLP. In natural language processing, useless words (data), are referred to as stop words. A stop word is a commonly used word (such as “the”, “a”, “an”, “in”) that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query.

Algorithm:

1. Input
- 8
2. if words in sentence == stopwords list then goto step-4
3. else message(“No stopwords”) then goto step-4
- 4.output
- 5.Exit

Stop Words List
The
An
In
On
There
Here
This
That
For
From
Who
Are

Fig 2.Stop words table

Input :

Syria Daesh claims a suicide attack in the Al-Zahra area of Homs. Three of the IS suicide bombers who attacked the military hotel in Aden are from Yemen.

Output :

Syria Daesh claims “a” suicide attack “in” “the” Al-Zahra area “of” Homs. Three “of” “the” IS suicide bombers “who” attacked “the” military hotel “in” Aden “are” “from” Yemen.

C. Filtration

Filtration helps remove all the punctuation marks as well as stop words from the text. After the application of filtration technique only the important text remains.

Algorithm:-

1. Input
2. if words in sentence == Filtration list then goto step-4
3. else message(“No filtration is present”) then goto step-4
- 4.output
- 5.Exit

Input :

#Syria Daesh claims a suicide attack in the Al-Zahra area of #Homs. Three of the IS suicide bombers who attacked the military hotel in Aden are from #Yemen.

Output :

Syria Daesh claims suicide attack AlZahra area Homs
Three IS suicide bombers attacked military hotel
Aden Yemen.

D. Stemming using Dawson’s technique:

Stemming is the process of reducing inflected (or sometimes derived) words to their word stem, base or root form—generally a written word form.

Why Stemming:

→ The suffixes changes meaning of the term even main root word is same.

→The ambiguity will be generated if suffixes are present.

→ Suffixes makes data complex and occupy memory.

What we achieve if we remove suffixes

→Ambiguity will be reduced.

→Meaningful words will be generated.

→ Root/stem word is formed.

→ Term Frequency values of word can be calculated correctly.

→ Reduce number of terms in the document for effective IR.

→Reduce size of data in the system to save memory utilization.

Algorithm:-

- 1.Input
2. if words suffix in sentence == suffix list then goto step-4
3. else message("No stopwords") then goto step-4
- 4.output
- 5.Exit

Input :

Syria Daesh claims a suicide attack in the Al-Zahra area of Homs. Three of the IS suicide bombers who attacked the military hotel in Aden are from Yemen.

Output :

Claims →claim

E. Morphology using Morphological parsing :

Morphological analysis may be defined as the process of obtaining grammatical information from tokens, given their suffix information. It analyzes a given token and generates morphological information, such as gender, number, class, and so on, as an output.

Algorithm:-

- 1.Input
2. if words suffix in sentence == suffix list then goto step-3
3. Separate the suffix from the word and pass on the remainder to step 4
4. Perform dictionary comparison with remainder else restore the word
- 4.output
- 5.Exit

Input:

Syria Daesh claims a suicide attack in the Al-Zahra area of Homs. Three of the IS suicide bombers who attacked military hotel in Aden are from Yemen.

Output :

Word : bombers

Suffix : ers

Remainder : bomb

Original word : bomb

III .IMPLEMENTATION DETAILS

Techniques

A.URL detection

Malicious URL, a.k.a. malicious website, is a common and serious threat to cybersecurity. Malicious URLs host unsolicited content (spam, phishing, drive-by exploits, etc.) and lure unsuspecting users to become victims of scams (monetary loss, theft of private information, and malware installation), and cause losses of billions of dollars every year. To improve the generality of malicious URL detectors, machine learning techniques have been explored with increasing attention in recent years. To improve the blacklists, machine learning techniques are constantly used to update them.

Semantic Analysis

Semantic analysis describes the process of understanding natural language—the way that humans communicate—based on meaning and context. The semantic analysis of natural language content starts by reading all of the words in content to capture the real meaning of any text. It analyzes context in the tweet that have more than one definition. It also understands the relationships between different concepts in the text. For example, it understands that a text is about "politics" and "economics" even if it doesn't contain the actual words but related concepts such as "election," "Democrat," "speaker of the house," or "budget," "tax" or "inflation."

B. n-gram Algorithm

An n-gram is a contiguous sequence of n items from a given sample of text or speech. An n-gram model is a type of probabilistic language model for predicting the next item in such a sequence in the form of a (n - 1) order. n-gram models are now widely used in probability, communication theory, statistical natural language processing, computational biology and data compression. Two benefits of n-gram models are simplicity and scalability – with larger n, a model can store more context with a well-understood space-time tradeoff, enabling small experiments to scale up efficiently. n-gram models are widely used in statistical natural language processing. In speech recognition, phonemes and sequences of phonemes are modeled using a n-gram distribution. For parsing, words are modeled such that each n-gram is composed of n words. For language identification, sequences of characters/graphemes (e.g., letters of the alphabet) are modeled for different languages. For sequences of characters, the 3-grams (sometimes referred to as "trigrams") that can be generated from "good morning" are "goo", "ood", "od ", "d m", " mo", "mor" and so forth, counting the space

character as a gram (sometimes the beginning and end of a text are modeled explicitly, adding "__g", "_go", "ng_", and "g__").

N-grams of texts are extensively used in text mining and natural language processing tasks. They are basically a set of co-occurring words within a given window and when computing the n-grams you typically move one word forward. For example, for the sentence "The cow jumps over the moon". If N=2 (known as bigrams), then the ngrams would be:

- the cow
- cow jumps
- jumps over
- over the
- the moon

IV. SAMPLE DATASET

An experiment is conducted in order to identify the input/output behavior of the system. Identify inputs. Specify the sample inputs that would be used in the experiments. The sample dataset used in the experiment are identified and given in Table 3.1

Name	User name	Tweet id	Time	tweets
Guns and Coffee	Guns and Coffee	2	01-06-2015 21:07:09	ENGLISH TRANSLATION: A MESSAGE TO THE TRUTHFUL IN SYRIA - SHEIKH ABU MUHAMMED AL MAQDISI: https://justpaste.it/MESSAGETOTHE TRUTHFULINSYRIA https://twitter.com/account/suspended

Fig 3. Sample Dataset

V. APPLICATIONS

There are various applications of this domain system. The applications are listed here

1. To safeguard nation's integrity

National Security is of prime importance for any nation to maintain peace and harmony. Nations face numerous internal security challenges and Social Media act as the platform for that. Social media is not security threat in itself but the users of these services can pose the threats by their anti-social endeavors. The biggest challenge for internal security of nation through social networking site is cyber terrorism. Today terrorists select Social Media as a practical alternative to disturb the function of nations and other business activities because this technique has potential to cause huge damage. It poses enormous threat in international system and attracts the mass media, the security community, and the information technology corporation. This system will help to reduce terrorism spread around the world and thereby ensure nation's security and integrity.

2. Facilitating intelligence agencies

This system proves helpful for antiterrorism departments/agencies such as cybercrime departments by

providing them the necessary terrorist details. Cyberterrorism is the intentional use of computers, networks, and public internet to cause destruction and harm for personal objectives. Experienced cyberterrorists, who are very skilled in terms of hacking can cause massive damage to government systems, hospital records, and national security programs, which might leave a country, community or organization in turmoil and in fear of further attacks. The objectives of such terrorists may be political or ideological. If a user is found spreading terror by posting any type of terror related media like tweets, images, videos, etc. then that user's details could be extracted which would eventually facilitate cybercrime department for further investigation.

3. Safeguard social sites

With the large population involved in social networks, this causes many issues that should be avoided and safeguarded against. Social media enables an individual or agency to communicate interactively and enables exchange of user generated content and it is explained by a number of tools, which includes blogs, twitter and social networking sites. The advantages of Social media are so many but they are posing threat to Internal Security in various forms like Cyber Terrorism, Fraud, crime, spreading violence, etc. As Internet is growing explosively, online criminals try to present fraudulent plans in many ways. Social networking sites also pose major challenge in financial and organized crime which destabilizes the system. It creates threat to a company's security because of what employees might disclose and they are on prime target for cyber criminals. The people who carry out terror related activities on social media sites can be reported and thus can be prevented from spreading terror information which results in safeguarding social sites.

VI. CONCLUSION

The information present on the web is generally in unstructured or semi structured format (more than 80%) such as email contents, HTML, XML, MP3, Videos etc. The text mining is tool which acts as Text Data Mining to discover the knowledge form the large volume of unstructured text without disturbing overall goal of searching. Text mining plays important role in the field of Data mining, information retrieval, machine learning, knowledge extraction systems etc. in this paper a literature survey on the techniques .The techniques such as Information Extraction, Summarization, Categorization, Topic tracking, clustering, EART are discussed in this paper. The Text mining tools can be applicable in many areas such as newspapers, media, health, insurance, market analysis, junk mails and many more.

The text mining is proved to have high commercial potential value. Companies mostly store their information in the text i.e. in unstructured format thus to retrieve meaningful information or to generate knowledge discovery, text mining plays a crucial role.

REFERENCES

[1] International Journal of Engineering Technology Science and Research IJETSRSR www.ijetsrsr.com ISSN 2394 – 3386 Volume 5, Issue,2018 (EXR:A proposed framework to detect potential suspects involved in Illicit activities via OSN)

[2] Real-time prediction of drive by download attacks on twitter,19 AUG 2017

[3] Real-time Classification of Malicious URLs on Twitter using Machine Activity, Data, 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Pete Burnap, Amir Javed, Omer F. Rana, Malik S. Awan, School of Computer Science and Informatics Cardiff University, Cardiff, UK

[4] A Survey on Identification and Analysis of Poor Quality Content on Facebook. Prateek Dewan, Indraprastha Institute of Information Technology - Delhi (IIITD), Comprehensive Examination Survey Report, Vol. 1, No. 1, Article 1, Publication date: October 2014.

[5] Online Social Networks and Terrorism 2.0 in Developing Countries, December 2013, Vol 1,no-4,ISSN-2345-3397, Fredrick Romanus Ishengoma, College of Informatics and Virtual, Education, The University of Dodoma, Dodoma, Tanzania.

[6] https://www.unodc.org/documents/frontpage/Use_of_Internet_for_Terrorist_Purposes.pdf

[7] <http://www.insiktintelligence.com/red-alert-system-for-online-terrorist-content>

[8] http://www.ise.bgu.ac.il/faculty/mlast/papers/jiw_paper.pdf

Mobile App for Stress Detection and Mental Health

Shruti Pawar, Amruta Salvi, Shifa Khan, and Neha Gawand

Dr. Madhumita Chatterjee

Department of Information Technology, PCE, Navi Mumbai, India – 410206

Abstract—In today’s society, working environments are becoming more stressful. Stress is a mental condition that everybody experiences in his life, sometimes even daily. As World Health Organization (WHO) says, Stress is a mental health problem affecting the life of one in four citizens. Human stress leads to mental as well as socio-fiscal problems, lack of clarity in work, poor working relationship, depression and finally commitment of suicide in severe cases. Automatic detection of stress minimizes the risk of health issues and improves the welfare of the society. This paves the way for the necessity of a scientific tool, which uses physiological signals thereby automating the detection of stress levels in individuals. Stress management systems play a significant role to detect the stress levels which disrupts our social economic lifestyle.

I. INTRODUCTION

In today’s society, working environments are becoming more stressful. Stress is a mental condition that everybody experiences in his life, sometimes even daily. As World Health Organization (WHO) says, Stress is a mental health problem affecting the life of one in four citizens. Human stress leads to mental as well as socio-fiscal problems, lack of clarity in work, poor working relationship, depression and finally commitment of suicide in severe cases.

Automatic detection of stress minimizes the risk of health issues and improves the welfare of the society. This paves the way for the necessity of a scientific tool, which uses physiological signals thereby automating the detection of stress levels in individuals. Stress management systems play a significant role to detect the stress levels which disrupts our social economic lifestyle.

The scope of this project is very vast in today’s life as the competition has arisen in various fields of technology. Everyone is in the race for giving their excellence and commonly tend to exhaust mentally and physically. In such situations, the mobile application will come in handy to detect levels of stress of the users and to alert their colleagues about their state.

The application will be implemented in Android Platform.

The project will consist of usage of the existing mobile sensors to sense the heart rate, footsteps will be counted as soon as the

user will take the phone and start walking, call logs that, include call duration and number of calls, GPS for sensing the current user location. The sensors will sense for any abnormalities and if found, an alert message will be sent to the guardians or colleagues of the user. Algorithms will be implemented as mentioned in further chapters.

II. LITERATURE SURVEY

1. Thomas Kowar, et al., June 2015, 'Smartphone Based Stress Prediction'

This paper ('Smartphone Based Stress Prediction') is related to predicting stress on the basis of questionnaire and usage of smartphone. The stress will be detected by the behavior of the user on the basis of the answers given by the users. The users need not expose their names. In their research they used an Activity Sensor which is available at Google Play store. The user session is taken into consideration. The session starts when the smartphone screen is turned on and session ends when the screen is turned off. Results show significant correlations between stress and smartphone data and outperform previously reported significance levels.

2. Mariana Kaiseler, et al., 2008, 'A Mobile Sensing Approach to Stress Detection and Memory Activation for Public Bus Drivers'

This paper ('A Mobile Sensing Approach to Stress Detection and Memory Activation for Public Bus Drivers') gives a description about the detecting stress using smartphones, Vital Jacket, disposable electrodes, GPS receiver, NetBook PC for bus drivers. The netbook is the device which will detect stress with the help of true sensors present in the VitaJacket. The Vital Jacket and the Netbook will be connected to each other via Bluetooth. The GPS receiver used was a Bluemax Bluetooth device that was placed near a bus window and transmits GPS information to the netbook via Bluetooth. The processing of the ECG signal was performed using the open-source library PhysioToolkit from Physionet. This was performed on actual

Bus drivers and the results showed that the methodology is successful in detecting stressful events based on bus driver's physiologic responses.

3. Ulrich Reimer, et al., 2017, 'Mobile Stress Recognition and Relaxation Support with Smart Coping: User-Adaptive Interpretation of Physiological Stress Parameters'

This paper ('Mobile Stress Recognition and Relaxation Support with Smart Coping: User-Adaptive Interpretation of Physiological Stress Parameters') describes a mobile solution for the early recognition and management of stress based on continuous monitoring of heart rate variability (HRV) and contextual data (activity, location, etc.). A central contribution is the automatic calibration of measured HRV values to perceived stress levels during an initial learning phase where the user provides feedback when prompted by the system. This is crucial as HRV varies greatly among people. A data mining component identifies recurrent stress situations so that people can develop appropriate stress avoidance and coping strategies. A biofeedback component based on breathing exercises helps users relax. The solution is being tested by healthy volunteers before conducting a clinical study with patients after alcohol detoxification.

4. Martin Gjoreski, et al., 2016, 'Continuous Stress Detection Using a Wrist Device – In Laboratory and Real Life'

This paper ('Continuous Stress Detection Using a Wrist Device – In Laboratory and Real Life') is related to the method for continuous detection of stressful events using data provided from a commercial wrist device. The method consists of three machine-learning. It consists of a laboratory stress detector that detects short-term stress every 2 minutes; an activity recognizer that continuously recognizes user's activity and thus provides context information; and a context-based stress detector that gives the output of the laboratory stress detector and the user's context in order to provide the final decision on 20 minutes interval. The method was evaluated in a laboratory and a real-life setting. The method is currently being integrated in a smartphone application for managing mental health and well-being. Even though the results show that there is still room for improvement, they are encouraging for such a challenging problem.

5. Sunghyun Yoon, et al., 2016, 'A Flexible and Wearable Human Stress Monitoring Patch'

A human stress monitoring patch integrates three sensors of skin temperature, skin conductance, and pulse wave in the size of stamp (25mm×15mm×72µm) in order to enhance wearing comfort with small skin contact area and high flexibility. The skin contact area is minimized through the invention of an

integrated multi-layer structure and the associated micro fabrication process; thus being reduced to 1/125 of that of the conventional single-layer multiple sensors. The patch flexibility is increased mainly by the development of flexible pulsewave sensor, made of a flexible piezoelectric membrane supported by a perforated polyimide membrane. In the human physiological range, the fabricated stress patch measures skin temperature with the sensitivity of $0.31\Omega/^{\circ}\text{C}$, skin conductance with the sensitivity of $0.28\mu\text{V}/0.02\mu\text{S}$, and pulse wave with the response time of 70msec. The skin-attachable stress patch, capable to detect multimodal bio-signals, shows potential for application to wearable emotion monitoring.

6. Jacqueline Wijsman, et al., January 2010, 'Trapezius Muscle EMG as Predictor of Mental Stress'

The ability to measure stress with a wireless system would be useful in the prevention of stress-related health problems. The aim of this experiment was to derive stress levels of subjects from electromyography (EMG) signals of the upper trapezius muscle. Two new stress tests were designed for this study, which aimed at creating circumstances that are similar to work stress. An experiment is described in which EMG signals of the upper trapezius muscle were measured during three different stressful situations. Stress tests included a calculation task (the Norinder test), a logical puzzle task and a memory task, of which the last two were newly designed.

7. Jesus Minguillion, et al., 2018, 'Portable System for Real Time Detection of Stress Level'

In this paper, they proposed a portable system for real-time detection of stress based on multiple bio signals such as electroencephalography, electrocardiography, electromyography, and galvanic skin response. In order to validate our system, we conducted a study using a previously published and well-established methodology. In our study, ten subjects were stressed and then relaxed while their bio signals were simultaneously recorded with the portable system. The results show that our system can classify three levels of stress (stress, relax, and neutral) with a resolution of a few seconds and 86% accuracy. This suggests that the proposed system could have a relevant impact on people's lives. It can be used to prevent stress episodes in many situations of everyday life such as work, school, and home.

8. Mandeep Singh, et al., December 2013, 'A Novel Method of Stress Detection using Physiological Measurements of Automobile Drivers' Stress while driving is an important factor in many numbers of fatal road accidents worldwide. There has been much work done in driver stress detection. In this research, we present a method based on a correlation analysis and developed a mathematical function for the estimation of

automobile driver stress level. The proposed methodology monitors driver's stress level using features extracted from selected physiological parameters. The results obtained indicate a strong correlation between the stress level of driver and the stress function formed. Threshold approach is used to perform a classification of effective states as "Low Stress", "Moderate Stress" and "High Stress" based on different traffic conditions. The stress function acts as a direct indicator of stress level of the automobile diver whose physiological parameters are monitored continuously under variable traffic conditions.

III. PROPOSED SYSTEM ARCHITECTURE

The system overview is presented in Figure 1

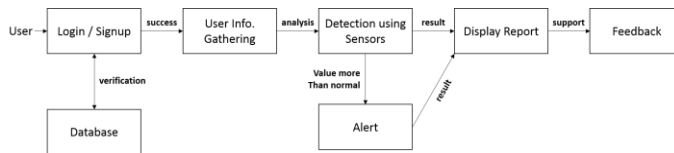


Fig. 1. Overview of Mobile App

This system helps user to understand the overview of the system used for stress detection. Based on this, recommendation techniques will have great influence throughout the project.

The following steps will be followed:-

Step 1: User will sign up or login into the application using his credentials. If they are found correct, user is provided further access.

Step 2: User's physical characteristics will be gathered in order to identify the correct states or levels for him.

Step 3: Sensors will detect for any abnormal behavior or levels of the user. If yes, then he will be alerted.

Step 4: Reports are displayed stating whether the user is stressed.

Step 5: Feedback and additional details if to be added.

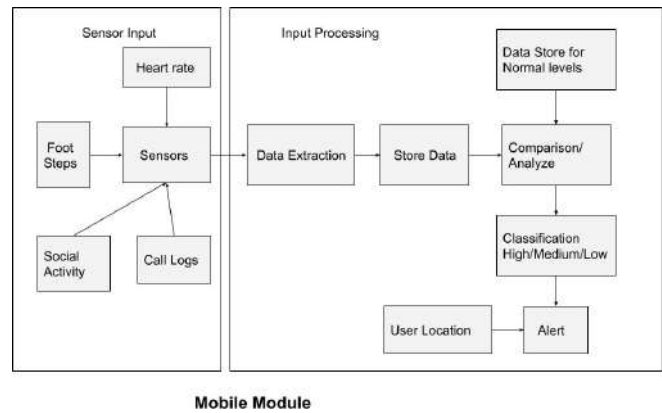
SENSOR INPUT:-

1. Heart Rate Sensor:

- This sensor will take heart beats as the input.
- The user will have to place a finger on the screen.
- After the pulses are counted they are stored into the database.

2. Footsteps:

- Footsteps will be counted as soon as the user takes the phone and starts walking.
- The count of the footsteps will be taken and stored in the database.



Mobile Module

Fig. 2. Proposed System Architecture

3. Messages/Social Activity:

- The messages that the user receives will be monitored.
- If the messages contain certain negative words or stress related words or stress related words then those messages will be stored.
- Depending upon the number of such words, the stress levels will be detected.

4. Call Logs:

- Call logs will be tracked on the basis of the number of calls and the duration of calls.
- 1) The monitoring time of call logs would be 1 hour.
- 2) If the user is receiving a lot of calls in 1 hour or speaking frequently for hours together, then it may be counted as stress causing factor.

INPUT PROCESSING:-

1. Data Extraction:

- All the data gathered from the sensors will be extracted.
- This extracted data will be stored in a database.
- This data will be later used for comparison.

2. Database for Sensed Values:

- This database will be store the sensed data.

3. Database for Normal Values:

- This is a database that will store the characteristics' normal values.
- The sensed characteristics will be compared to the normal values.
- Based on the above observation, the results will be displayed later.

4. Comparison/Analyze:

- This module will receive from two sources i.e., database for sensed values and database for normal values.
- The next step will be analyzing stress levels.
- Analyzing will be analyzing stress level.

- Depending upon this computation the stress levels will be displayed.

5. Classification:

- This will receive input from Comparison/Analyze module.
- If the values are not matching with normal levels then this module will classify stress level into 'High', 'Medium', & 'Low'.
- The stress level will go 'High' if more than 3 parameter values are not normal.

6. User Location:

- This module will be used while sending alert to the related guardian.
- While informing the guardian, the user location will also be sent.

7. Alert:

- This module will receive inputs from 'User Location' and 'Classification' module.
- The alert module will send an alert message to the respective guardian automatically when the stress level is high.
- The user's location will also be sent.

IV. REQUIREMENT ANALYSIS

Following will be the hardware and software requirements:-

Table 1: Hardware Details

System	Intel i7 8gen 3.2GHz
Hard Disk	100 GB SSD
Mobile Phone	Android
Ram	4 to 8 GB

Table 2: Software Details

Operating System	Windows 7 & above
Coding Language	Java
Front End	HTML
Back End	MySQL

V. SUMMARY

The project involves building a mobile application for detecting stress levels and tracking mental health of students. This application uses all the sensors built into mobile devices to measure physical changes. It should record environmental noise and tracks calls and text messages. There can be over many values the application can record, including surrounding

noise level; social activity, as monitored by texts and calls; changing environmental conditions, measured through air pressure as well as light level; and even posture, measured by the phone's accelerometer.

You can also track moods expressed through emoticons and use attached monitors to provide pulse and heart-rate data. Our goal is to create and combine a continuous monitoring device and stress management device into one system (Application). Our continuous monitoring device (application) will be responsible for monitoring the users stress level, so that the user will be able to concentrate on his/her tasks throughout the day and be assured that stress levels are accounted for. We will also help the user regulate his breathing to relieve any stress that is detected.

ACKNOWLEDGMENT

We are profoundly grateful to Dr. Madhumita Chatterjee for her expert guidance and continuous encouragement throughout to see that this project rights its target since its commencement to its completion.

We would like to further thank the HOD of IT Dr. Sharvari Govilkar, the faculty for providing us to pursue this project and also in helping us to guide and decide between plethora of options the college had to offer.

We would like to thank Dr. Sandeep Joshi providing the required resources and guidance for the project, without his support this project would not have been possible, and we are grateful for his encouragement.

REFERENCES

- [1] Henner Gimpel. 'myStress: Unobtrusive Smartphone-based Stress Detection', 2015. [Online].
- [2] Michel Deriaz. 'Stress Detection Using Smartphone Data', 2016. [Online].
- [3] Panagiotis Kostopoulos. 'Stress Detection Using Smartphone Data', 2013. [Online]
- [4] Jon White, 'Smartphone App for sensing stress airs monitors environment and physiological', March 2013. [Online]
- [5] Jacqueline Wijsman, 'Trapezius Muscle EMG as Predictor of Mental Stress', 2010. [Online]
- [6] Jesus Minguillion, August 2018, 'Portable System for Real-Time Detection of Stress Level'
- [7] T. Fohr, A. Tolvanen, T. Myllymäki, et al., "Subjective stress, objective heart rate variability-based stress, and recovery on workdays among over weight and psychologically distressed individuals: A cross-sectional study", Journal of Occupational Medicine and Toxicology, vol. 10, no.1, p. 39, 2015.

- [8] M. Adam, H. Gimpel, A. Maedche, and R. Riedl, N“Stress-sensitive adaptive enterprise systems: Theoretical foundations and design blueprint”, in Gmunden Retreat on NeuroIS 2014 Proc., F. Davis, R. Riedl, J. vom Brocke, et al., Eds., Gmunden, Austria, 2014.
- [9] A. Muaremi, B. Arnrich, and G. Troster, “Towards measuring stress with smartphones and wearable devices during workday and sleep”, *BioNanoScience*, vol. 3, no. 2, pp. 172–183, 2013.
- [10] M. Morris and F. Guilak, “Mobile heart health: Project highlight”, *IEEE Pervasive Computing*, vol. 8, no. 2, pp. 57–61, 2009

Sentiment Analysis Based on Comments from Online Social Network

Vedant Patil, Jayesh Thakur, Kapildev Yadav and Prof. Deepti Lawand

Department of Information Technology, PCE, Navi Mumbai, India - 410206

Abstract— Internet is the platform where most of us share our happiness or other feelings. Recent years are devoted in studying and mining the data which is on social platform. This task includes understanding explicit and implicit information conveyed by sentiments. It can be extracted from the comments on social media using dictionary-based sentiment analysis or Review-Seer. Comments of the person are important to analyze the sentiments of the person at the time of writing the comment. The task is to classify the comments into positive, negative and neutral sentiments further into different emotions, for which it uses the concept of Plutchik's wheel of emotions and further makes a dictionary. The system will take input from user to classify and predict the emotions and strength of that emotion (Negative Emotions). There are basic eight emotions and system will primarily focus on negative emotions. Plutchik's wheel of emotion gives joy and sadness, anger and fear, trust and disgust, surprise and anticipation. The use of Plutchik's wheel of emotions will provide the real emotional view of comments. The confidence of the will be given which will indicate the strength of feeling. It uses fuzzy logic approach using Naïve Bayes or decision tree algorithm for prediction and generates output.

Keywords—Sentiment Analysis, Plutchik's Wheel, Machine Learning, Data Mining.

1. Introduction

Recently there has been a growing interest in social media and using it to update lifestyle. These comments entered by user contains pure emotions that needs to be extracted using different data mining algorithms. The task of mining sentiments and opinions from natural language is difficult one. It involves an intense understanding of most of the implicit and explicit information which is conveyed by structure of language. The availability of a dynamic corpus contains the user generated data, such as reviews for products or polling data. Big data is the large amount of easily available data on web, Social media, remote sensing data, etc. in form of structured data, semi-structured or unstructured data. We can use this large data for sentiment analysis. Sentiment analysis is the opinion mining used on the web for identifying the text. It is nothing but to get the real voice of people for specific product, services, movies, news, issues from online social networking site like Twitter. This data contains many important aspects which will be helpful in judging the turn of tide in market trend.

2. Literature Survey

A. Fine-grained Sentiment Analysis with 32 Dimensions[1].: This system does deal with range of total 32 emotions. It uses concept of Plutchik's wheel of emotion to classify comments into different 32 sentiments. The mathematical model of Naïve Bayes is used for classification and prediction uses intensity-based technique. Using Naïve Bayes classifier to classify into 32 emotions also makes increase in accuracy. The eight basic emotions given is used as classes to classify emotions into.

B. Rule-based Emotion Detection on Social Media: Putting Tweets on Plutchik's Wheel[2]: This paper studies and analyze sentiments beyond polarity and uses Plutchik's wheel of emotions. It uses extension of Rule based emission model. This model thinks beyond the normal metrics of sentiment analysis using polarity and uses Rule-Based Emission Model (RBEM) algorithm (Tromp and Pechenizkiy, 2013) that can be used for polarity detection assigning new messages a label that is one of positive, neutral, negative. Important in algorithm is that positivity and negativity are opposite and allows negation to simply invert the emission. RBEM uses pattern matching and uses wildcards for it. The model used is compact as well as complete which works well with RBEM-Emo which is stated as extension of Rule Based detection algorithm.

C. Combining Strengths, Emotions and Polarities for Boosting Twitter Sentiment Analysis[3]: This paper proposes an approach for boosting twitter sentiment classification using different sentiment dimensions as meta-level features. This research shows the combination of sentiments improves the twitter sentiment classification tasks. The scopes of tweets are categorized upon some categories as polarity, emotion, strength. It does different testing with different types of algorithm. It uses classification approach like OpinionFinder Lexicon, AFFIN Lexicon, SentiWordNet Lexicon, SentiStrenght Lexicon, Senti140 method, NRC Lexicon. So, when it

classifies tweets into polarity classes, we are essentially projecting these multiple dimensions to one single categorical dimension. But also, sentiment classification of tweets can lead to loss of valuable sentiment information.

D. Sentiment Analysis on Product Review Using Plutchik Wheel of Emotion with fuzzy logic[4]: The model consists of stemming and stopword techniques. This filtering removes almost all unwanted noise from comment. The filtered comment is then split to get the separate words for comparing. Then each single word is compared with the sentiment words dictionary. If the word is matched with the positive or negative dictionary then it is placed in the corresponding box, that is positive word in positive words text and in the same way negative words are placed. The comparison is done between number of positive word and number of negative words in a given comment. The condition is checked whether the positive words are more or negative and accordingly the comment is decided to be positive or negative. If both the positive and negative words are same or if there are no positive or negative, the comment is treated as neutral comment.

2.1 Summary of Related Work

Sentiment Analysis has been largely developed in the recent years due to the requirement of sentiments of people. Due to this, a large number of importance has been given to neural networks, machine learning, etc. For ex, deep neural network classifiers has been proposed in [1] where in the model proposed is including all the 32 emotions and has used the naïve bayes technique which we implementing in our model. The emotions are further put on the Plutchik's and the emotion is derived. Existing.

System Has Following Steps:

Data collection using twitter API: Publicly large sets of Twitter data are not available. Hence, they first extracted twitter data through twitter API.

Data Preprocessing[4]: It involved cleaning of data by spell correction punctuation etc. Reducing noise from the data.

Applying Classification Algorithm: The Classification Algorithm is applied on tweets to categories them with highest accuracy.

Classified tweets and result: The tweets are further classified three defined categories. Result of which is displayed in form of pie chart.

One of the techniques used in sentiment analysis is Rule Based Emotion Detection (RBEM) [2]. The RBEM thus efficiently increases the exactness of the developed model where the accuracy is guaranteed and is tightly coupled with the plutchik's wheel of emotion. The data collection,

preprocessing, clustering, sentiment classification, prediction[4] are various processes that are been applied to get accurate output with the help of plutchik's wheel of emotion.

3. Proposed Work

The proposed system will primarily classify emotion into positive, negative, neutral and further into 8 basic emotions. Firstly, the data required for analysis will be divided into testing and training datasets, this dataset are downloaded from official twitter APIs. The Naive Bayes classifier will be trained according to this dataset.

3.1 System Architecture

The system architecture is given in Figure 1.

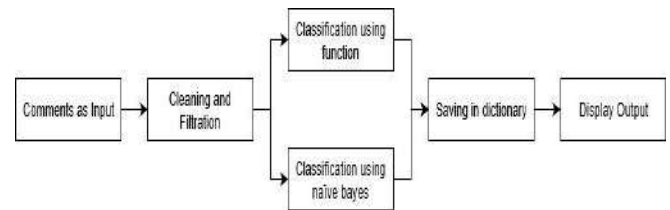


Fig. 1 Proposed system architecture

A. Comments as Input: The system will accept textual inputs which are in comments format. These comments will be entered by user after signing in on our website and typing comments on their page. These comments will be further cleaned to analyze them.

B. Cleaning And Filtration: The second part receives the input comments entered by user. This comment cannot be directly used to analyze the emotions as it contains some amount of excessive information which is not use full for sentiment extraction, such as Nouns, Names of places, etc. So we will be using extraction using predefined tokenizer known as TextBlob. It will extract the noun and sentiment defining word. Textblob can tag the word with part of speech which can be further useful while training naïve bayes classifier. Other processes are included like lemmatization.

C. Classification using Predefined Function: After cleaning input comment it is further used for extracting basic emotion. We will be using predefined function for extracting basic classification into positive negative and neutral. The classifier will be trained on training dataset.

D. Classification using naïve bayes: Naive Bayes classifier will train using previous training and test

datasets. This naive Bayes classifies the comments into 8 classes named as anticipation, joy, anger, sadness, surprise, disgust, fear and trust. This comment will be saved along with its tag of class into sentiment dictionary for further learning of system. We can train classifier with:

```
cl = NaiveBayesClassifier(train).
```

We can also find accuracy of test set using:

```
cl.accuracy(test)
```

After training output which has to be given using naive Bayes classification technique. We know the formula which can be used for classification as,

$$P(A|B) = P(B|A) * P(A) / P(B)$$

Where, $P(A|B)$ is probability the A belongs to class B, $P(B|A)$ is evidence, $P(A)$ is probability of class A is seen and similar with B.

The $P(A|B)$ is calculated for each word and then the class tag is selected with maximum probability. Maximum probability is selected and saved into dictionary for further increasing accuracy of classifier.

E. Saving to dictionary: After classification of comment into three basic emotion and then prediction into one of the eight emotion, comment is further saved into database for future predictions. The comment is saved along with tag of the emotion and further prediction will be done.

E. Output Prediction: After saving into dictionary, output will be displayed to user. Output will give the basic classified emotion of that comment. It will give positive, negative or neutral along with polarity of comment. Also it will give the prediction of emotion of comments from eight basic emotion given in the plutchik's wheel of emotion.

3 Requirement Analysis

The implementation detail is given in this section.

3.1 Software

Software requirements are Html and Bootstrap supporting browser, as the systems user interface is based on that versions. Also Flask framework is used for classifier.

3.2 Hardware

Strong internet connection is recommended for website access. Otherwise no hardware requirements are there.

3.3 Dataset and Parameters

^[1]Dataset used contains comment along with tags of positive, negative and neutral emotion. Used dataset mostly have no neutral comment as it doesn't have any effect on classifier performance.

ACKNOWLEDGMENT

It is our privilege to express our sincerest regards to our supervisor Prof. Deepti Lawand for the valuable inputs, able guidance, encouragement, whole-hearted cooperation and constructive criticism throughout the duration of this work. We deeply express our sincere thanks our Head of the Department Dr. Sharvari Govilkar and our Principal Dr. Sandeep M. Joshi for encouraging and allowing us to presenting this work.

REFERENCES

1. *Xianchao Wu, Hang Tong, Momo Klyen*, "Fine grained Sentiment Analysis with 32 Dimensions", A.I. & Research Microsoft Development Co. Ltd., Department of Mechano-Informatics, The University of Tokyo.
2. *Erik Tromp, Mykola Pechenizkiy*, "Rule-based Emotion Detection on Social Media: Putting Tweets on Plutchik's Wheel", The Netherlands, 15 Dec 2015.
3. *Felipe Bravo-Marquez, Marcelo Mendoza, Barbara Poblete*, "Combining strength, emotions and polarities for boosting Twitter sentiment Analysis", Chile, 11 Aug 2013.
4. *Dhanshri Chafale, Amit Pimpalkar*, "Sentiment Analysis on Product reviews using Plutchik's Wheel of Emotion with Fuzzy Logic", Nagpur University, Dec 2015
5. UCI Machine Learning Repository: Sentiment Labelled Sentences Data Set. Available: <https://archive.ics.uci.edu/ml/datasets/Sentiment-Labelled-Sentences> [Accessed:21Aug-2018].

Text Document Clustering Using Latent Semantics Indexing

Madhu Nashipudimath*, Ankita More†, Ameya Pokharkar†, Aditya Sawant†, Mayur Walshinge†

*Assistant Professor, Computer department Pce New Panvel

†Students, Department of Information Technology, Pce, New Panvel

Email:- *madhumn@mes.ac.in, aarun31@student.mes.ac.in, adityans15it@student.mes.ac.in, walshingemal5ith@student.mes.ac.in, ameyasp15it@student.mes.c.in.

Abstract -Text Documentation is an important and basic form of information distribution in human life whereas, document clustering is an important tool to help managing the vast amount of digital text document. Present scenario involves traditional approach of document clustering which include some problems like polysemy, synonymy, ambiguity and semantic similarity. This problem may not be captured by traditional mining technique. Hence semantic clustering is proposed for developing cluster related to keywords. The proposed method for semantic clustering is carried as first pre-processing followed by indexing (using inverted index). Trimming (using TF-IDF Threshold for creation of document matrix), later latent semantic is used to extract important features from term document matrix. Seed selection is carried to identify the centroids for clusters via Pillar Algorithm. K-mean clustering is performed based on these seeds. A model is proposed using this techniques to perform document clustering.

Keywords-Clustering, Latent Semantics Indexing (LSI), K-mean clustering.

I. INTRODUCTION

Text document is an important and basic form of information distribution in human life. The rise of Internet and cloud technology boosts the spread of digital text document. This trend positively reduces the use of non-eco-friendly paper and makes document management like searching, editing, and privilege assessment easier. However, the massive growth of digital technology creates another difficulty in managing document. For example, retrieving a document from a huge repository is a challenging task. One solution to reduce complexity in searching a document is by employing a clustering technique, either by clustering documents before search process or grouping search results [1].

Text clustering is a technique for merging similar documents into a group. Apart from document searching, clustering has been intensively used for news recommendation [3], topic detection, and document

Summarization, and transforming unstructured document to structured one.

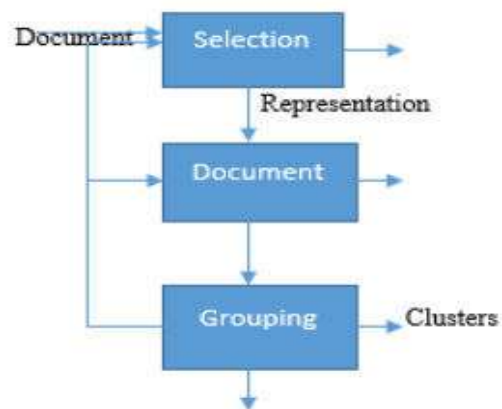


Fig1. Different stages in text clustering

A common feature used for clustering document is bag-of-words model. Possible feature is term frequency, relative term frequency, or tf-idf (term frequency and inverted document frequency) [8]. This simple model usually leads to a sparse vector since the dimensionality of document is huge. There is a large chance that the vector contains many zeros. This condition introduces a threat to many clustering methods which rely on similarity measure. An attempt to limit the number of features involved in clustering, it is divided into two categories: feature selection and feature extraction. The former selects a subset of existing features based on some principles while the latter transforms the features into other ones with lower dimensionality.

In the other hand, feature extraction are achieved by methods such as Latent Semantic Indexing (LSI) and Independent Component Analysis (ICA). LSI tries to reveal the most representative features from a document. Thus the dimension of tf-idf matrix is significantly reduced and then the clusters are developed. Thus with the help of LSI in combination with clustering method a new model is proposed for easy retrieval of information and presenting it to use.

II. RELATED WORK

In order to retrieve the information and present it to the end-user clustering is needed. It will be used to improve the ability to find the associated documents which are relevant to the given query in most simplified version. Research in this domain have used text clustering by enabling whole information results such that it can searched by the user easily, besides efficient way of browsing. By defining that text clustering is often an effective way to generate better results in a set of given documents [2].

Clustering is the task of organizing unlabelled objects in a way that objects in the same group are similar to each other and dissimilar to those in other groups. In other words, clustering is like unsupervised classification where the algorithm models the similarities instead of the boundaries. Clustering in text document is applied in order for easy retrieval of query from a huge repository manually this will require a huge amount of time in order to some time as well as human resource clustering provides a boon. Many clustering algorithm are developed in current situation which can handle various types of data among all clustering algorithms, K-means is a popular choice due to its simplicity. The objective of K-means algorithm is to partition data into K number of cluster by minimizing the distance of each data point into its centroid. In spite of its popularity and simplicity, K-means suffers from some problems. First, the number of clusters must be defined in advance [1]. Second, K-means is only effective for spherical data [1]. Third, K-means is sensitive to outliers [1]. The last problem arises from the nature of random seed selection in the first step of K-means. It means a unique clustering result is not guaranteed [1]. Therefore, K-means may not achieve the global optima, but a local One. Many approach has been identified to eliminate the disadvantages of K-means. One of them is Pillar algorithm [1]. The algorithm improves some drawbacks of K-Means by implementing deterministic seeds selection, outlier avoidance, singleton and empty cluster prevention. This paper, propose a framework for clustering text documents.

The results of clustering are improved by using the ontology based clustering algorithms rather than simple clustering algorithms [12] a large collection of text documents are considered as unstructured data. It is very difficult to group the text documents. A dataset is used for the clustering of documents. New framework to cluster text document has been introduced in this paper. It incorporates feature extraction using LSI and seed selection using Pillar algorithm. Experiment with newsgroup dataset show a performance improvement compared to classic K-Means [1].

Typical text clustering involves:

- Representation of document
- Similarity of document
- Grouping the document

A relative test for clustering involves two structures and measures and for external assessment it derives recovered structure and for internal validity it defines appropriate data [2]. In this research, we develop clustering algorithms which consists the unique features like sequential relationship, high performance of high end database, frequent word sequencing and word meanings. By all this unique functions, clustering algorithm attains better performance and high speed results at the same time than the other regular methods.

Latent semantic indexing, sometimes referred to as latent semantic analysis, is a mathematical method developed to improve the accuracy of information retrieval. It uses a technique called singular value decomposition to scan unstructured data within documents and identify relationships between the concepts contained therein. In essence, it finds the hidden (latent) relationships between words (semantics) in order to improve information understanding (indexing). It provided a significant step forward for the field of text comprehension as it accounted for the contextual nature of language. Latent semantic analysis assumes that there is some relationship between words, by using the method of Statistical to extract the implicit latent semantic relation, thus effectively and accurately represent text information [3]. Using the method of latent semantic analysis to structure semantic space, or by optimizing the latent semantic analysis method of faults, to improve the retrieval accuracy. But in the process of retrieval, the query vector should be carried out to do some similarity calculations with each text vector, it requires a lot of time, in order to solve this problem, this paper[3] put forward a kind of latent semantic information retrieval algorithm based on pre-clustering, adopts the method of pre-clustering to pre-process document collection, when retrieving, there is only similarity calculation between query vector and the clustering centre, so as to avoid the similarity calculation between query vector and each text vector respectively, and to improve the efficiency of retrieval.

Latent semantic indexing works with every set of document provides so in this paper we use newsgroup as a dataset for clustering Newsgroup 20 this dataset consist a collection of more than 2000 news set so it is easy for any person to search for a query based on the newsgroup Latent Semantic Analysis (LSA) for automatic clustering of news articles yields a good results [4] The clustering scheme used within this work is based on the use of LSA and it consists of three phases. In the first

phase, a term-document matrix is constructed and decomposed to a concept space using LSA. Next, the dimensionality of the concept space is reduced and, after that, hierarchical clustering is performed in the third phase.

Choosing the right algorithm for clustering is an important task although k-mean algorithm is popular and widely use but it too has some drawback in order to overcome those pillar k-mean is selected [6]. Every clustering algorithm greatly rely upon the correctness of the initial centroids. Usually the initial centroids for the K-means clustering are determined randomly so that the determined centroids may reach the nearest local minima, not the global optimum. New approach to optimizing the designation of initial centroids for K-means clustering. This approach is inspired by the thought process of determining a set of pillars' locations in order to make a stable house or building. We consider the pillars' placement which should be located as far as possible from each other to withstand against the pressure distribution of a roof, as identical to the number of centroids amongst the data distribution. Therefore, our proposed approach in this paper designates positions of initial centroids by using the farthest accumulated distance between them. First, the accumulated distance metric between all data points and their grand mean is created. The first initial centroid which has maximum accumulated distance metric is selected from the data points. The next initial centroids are designated by modifying the accumulated distance metric between each data point and all previous initial centroids, and then, a data point which has the maximum distance is selected as a new initial centroid. This iterative process is needed so that all the initial centroids are designated. This approach also has a mechanism to avoid outlier data being chosen as the initial centroids. The experimental results show effectiveness of the proposed algorithm for improving the clustering results of K-means clustering. It designates the initial centroids' positions by calculating the accumulated distance metric between each data point and all Previous centroids, and then selects data points which have the maximum distance as new initial centroids. Our algorithm distributes all initial centroids according to the maximum accumulated distance metric. The algorithm also introduces a mechanism for detecting outliers. Several experiments involving eight benchmark data sets with five validity measurements and execution time were conducted. The experimental results show that our proposed algorithm is able to optimize the selection of initial centroids and improve the K-means precision in all data sets and in most of the validity measurements. Moreover, our proposed algorithm outperformed the other algorithms in at least two validity measurements, and the other initial centroid optimization algorithms in at least three

validity measurements. However, inappropriate parameter set-up in the proposed algorithm for outlier detection may lead to reduced performance. Adjusting to the characteristics of the data distribution in the data set is needed in order to set-up the appropriate parameters for the outlier detection mechanism [6]. New framework to cluster text document has been introduced in this paper. It incorporates feature extraction using LSI and seed selection using Pillar algorithm. Experiment with newsgroup dataset show a performance improvement compared to classic K-Means [4]. Work should be done to investigate the impact of D on performance. An automated method to choose optimum D is favourable. Also, an approach to automatically select the best value of K in K-Means should be investigated [1].

III. ISSUES IN DOCUMENT CLUSTERING

Since we are dealing with document clustering choosing the right document is important aspect considering if the document is made available from individual machine or document is downloaded from open source possibility of redundant data in the document are vast in order to cluster the data some parameter must first be assigned.

Choosing the right clustering method is also an critical decision as k-mean clustering algorithm is the most widely and popular method which is currently used for clustering of document but it have some limitation as well as drawback in order to put an check clustering algorithm should be chosen wisely and as discussed almost all clustering technique requires centroid for seed selection, seed selection is base for clustering the words therefore choosing the centroid is important task.

In order to implement latent semantic indexing on the document unstructured data is the big problem LSI mainly concentrates on linear model. Deciding on the number of topics is based on heuristics

IV. SOLUTION RECOMMENDED

Following suggestions open scope for further research:

- In order to remove the redundant data from the document pre-processing is suggest. There are five process used in pre-processing. The first step is case folding which switch all letters into lowercase ones and then followed by numbers and punctuations removal. The third step of pre-processing is stop word removal, a process to discard words with no significant meaning. One publicly

available stop list is SMART stop word list [1]. After that, sentences are split into words. The last step is stemming, where a root word of an inflected word is extracted. Incorporating a stemmer reduces the number of words involved in clustering and discovers more relation of words. Porter algorithm [8] is a common method for English word stemming.

- Seed selection is a deterministic approach to select initial centroids (seeds) in pillar K-Means is conducted by spreading the seeds away as far as possible from each other. To do that, the mean of data distribution is calculated and distance of each data point to the mean is calculated
- Use of the latent semantic indexing while clustering the words make user reliable to use the document and find the query in the document provided as input
- Use of the latent semantic indexing while clustering the document make user reliable to use the document and find the query in the document provided as input.

V. CONCLUSION

The advances in technology of computers and electronics, the increasing popularity of the Internet And the World Wide Web have led to vast amounts of increase in electronic text information. In order to browse text databases and extract relevant information quickly, efficiently, and correctly organizing digital text documents automatically has become an important research issue.

Different alternatives for document representation are discussed and chosen to use the Vector Space Model representation, where each distinct stemmed word is defined as a term. Stop words are removed, words are stemmed by using Porter's stemming algorithm, and dimensionality is reduced by the Latent Semantic Indexing using Singular Value Decomposition technique. The goal of information retrieval is to make it easy for the user to obtain data relevant to user request automatically. The performance of the Vector Space Model depends on the term-weighting schemes. Some popular term-weighting schemes are compared and a few new term weighting schemes are proposed, facilitating an easier interpretation of the results than the cosine similarity. Whenever the size of the database increases, it is essential to cluster the collection and then the retrieval algorithm is performed. For clustering, three different clustering algorithms are used k-mean, k-medoids and pillar k-

mean amongst which it was found that pillar k-mean is best amongst but future advancement should be done.

REFERENCES

- [1]Sigit Adinugroho, Yuita arum sari, M.Ali Fauzi, Putra Pandu Adhikarap: Optimizing K-Means Text Document Clustering using Latent Semantic Indexing and Pillar Algorithm. [*International Symposium on Computational and Business Intelligence*] (2017).
- [2]Venkata Srikanth Reddy,Patrick Kinnicutt,Roger Lees: Text Document Clustering:The Application of Cluster Analysis to Textual Document. [*International Conference on Computational Science and Computational Intelligence*](2016)
- [3]Chen Wenli: Application Research on Latent Semantic Analysis for Information Retrieval. [*Eighth International Conference on Measuring Technology and Mechatronics Automation*](2016)
- [4]Michal Rott, Petr Cerva: Latent Semantic Analysis for Clustering of News Articles. (*25th International Workshop on Database and Expert Systems Application*)(2014)
- [5]Ashwini Deshmukh, Gayatri Hegde: A Literature Survey on Latent Semantic Indexing. (*International Journal of Engineering Inventions ISSN: 2278-7461, www.ijejournal.com Volume 1, Issue 4 (September 2012)*)
- [6]Ali Ridho Barakbah and Yasushi Kiyoki: A Pillar Algorithm for K-Means Optimization by Distance Maximization for Initial Centroid Designation [Member, *IEEE*] (2010).
- [7]D. D. Lewis, Y. Yang, T. G. Rose, and F. Li, "RCV1: A New Benchmark Collection for Text Categorization Research," *J Mach Learn Res*, vol. 5, pp. 361–397, Dec. 2004.
- [8]Jaime Arguello. INLS 509: Information Retrieval jarguell@email.unc.edu. February 13, 2013. Wednesday, February 2013.
- [9] J. Zobel and A. Moffat: Inverted Files for Text Search Engines. *ACM Comput Surv*, vol. 38, no. 2, Jul. 2006.
- [10] T. Hofmann: Probabilistic latent semantic indexing. In *Proceedings of SIGIR'99*, 2012.
- [11]Barbara Rosario: Latent Semantic Indexing: An overview *INFOSYS 240 Spring 2000*
- [12]TwinkleSvadas,Jasmin Jha:Document Cluster Mining on Text Documents. (*International Journal of Computer Science and Mobile Computing*), (June 2015).

Author Biographical Statements

	<p>Madhu Mahesh Nashipudimath, has done her BE in computer Science and Engineering from Karnataka University and her M E in Computer Engineering from Shivaji University - Kolhapur. She has experience of working in Engineering and polytechnic colleges. She is presently working as Assistant professor at Pillai College of Engineering New Panvel. Her areas of interest are Neural networks, fuzzy logic, Data mining, Big data and their application in various fields. She has more than twenty national papers and fifteen international papers in her credit. She also attended number of workshops and training. She is a life Member of ISTE. She has total 20 Years of teaching experience. Presently she is pursuing her Ph.D. in the domain of Big data analytics.</p>
	<p>Ankita More is a final year undergraduate student, Department of Information technology from pillia's college of engineering (Panvel), University of Mumbai. She was appreciated for her work on the project USB write blocker using Arduino Nano and Humidity Temperature (Third Year).</p>
	<p>Ameya Pokharkar is a final year undergraduate student, Department of Information technology from pillia's college of engineering (Panvel), University of Mumbai. He was appreciated for his work on the project Android application for Swatch Bharat Mission (Third Year).</p>
	<p>Aditya Sawant is a final year undergraduate student, Department of Information technology from pillia's college of engineering (Panvel), University of Mumbai. He was appreciated for his work on the project for making interactive quiz holding website (Third Year).</p>
	<p>Mayur Walshinge is a final year undergraduate student, Department of Information technology from pillia's college of engineering (Panvel), University of Mumbai. He was appreciated for his work on the project Android application for Swatch Bharat Mission (Third Year).</p>

Real Time Traffic Event Detection Using Tweet Stream

Mentor/Guide: Prof. Sagar Kulkarni

Mohit A Rai
Student,
Pillai College of Engineering

Sumod Menon
Student,
Pillai College of Engineering

Riya Sawant
Student,
Pillai College of Engineering

Devbrat Singh
Student,
Pillai College of Engineering

Abstract—Twitter generates around 6000 tweets every second. Thus, it has become real-time source to detect any events happening around us. On the other hand, alerts on events like road-accidents, traffic and landslide do not reach the concerned officials and people on time. At present, there is no concrete system to update occurrence of these events quickly. Although, it is possible to get road-traffic information through various systems, there is no system which provides information about the events causing them. Thus, we propose a system to provide real-time updates on accidents, traffic and landslide using Twitter. This will provide fast updates on these events to the people and notify concerned officials in real quick time. Hence, required preventive measures can be taken quickly. The goal of this system is to assign class label to tweets depending on which events they belong to. Initially we will use text mining, natural language processing and machine learning to retrieve information from raw tweet. For classification, we use SVM classifier. Apache Spark will be used as database as it is analytics database for big data processing, with built-in modules for streaming and Machine Learning.

Keywords: Twitter, Traffic events, Text mining, Natural language processing, SVM classifier, Apache Spark.

I. INTRODUCTION

NLP is a way for computers to analyze, understand, and derive meaning from human language in a smart and useful way. NLP is used to analyze text, allowing machines to understand how humans speak. NLP is commonly used for text mining, machine translation, and automated question answering. On the other hand, Social Networking sites have become integral part of our life. Variety of social networking sites have cropped up which includes Facebook, Twitter, Instagram as more popular ones. Along with the entertainment factor, these sites have become rich source of information such as news, events or natural calamities.

On Twitter, people share the events happening around them. Twitter is popular due to its limited words status update which is short and on point twitter has become quite popular. People continuously share events happening around them. Traffic police, Government of India, and even Airlines companies entertain complaints on twitter and respond quickly as everyone see these responses. Moreover, since these

updates can be seen by anyone, these updates tends to be more accurate Hence, Twitter has emerged as a big raw field for data analyst as various new patterns can be discovered through the data analysis. Using knowledges discovered due to, we can find the events that are causing the traffic. These events will give users the more detailed idea of the traffic and estimate the duration for which the traffic will last. Currently, users are notified the current traffic condition but no information about events causing them are provided.

II. SCOPE OF THE PROJECT

Real Time Traffic Event Detection System will provide updates on the events which are causing traffics on the roads. The system will collect tweets to detect these events. The tweets which will be fetched from twitter API is unstructured and hence need to be processed using various preprocessing techniques. Finally, the tweet will be classified into the traffic related events using SVM classifier algorithm. Currently, the system will focus on detection of two main events that is accident and landslide. Rest of the traffic related events like construction, water-logging etc will be classified as general traffic related tweets. The system currently focuses on the traffic data obtained from Twitter in the English language from Mumbai region only. Further, we are assuming currently that all the tweets to be authentic. The system will be integrable with Google Maps API which will help to display the current traffic scenario along with the events causing them.

III. RELATED WORK

Large number of research paper and information related to text mining, Twitter data analysis, preprocessing, and Classification algorithm are available on Web. Medha A. Shah and Ketan R. Pandhare in the year 2017 in [1] presented a paper which detects real traffic events. Twitter is used as a source of data retrieval social network and Apache Spark is used for real time data-streaming. The retrieved tweets are classified using two classification algorithm namely Logistic Regression and SVM. The goal of the system was to just classify the tweets using these two algorithms and compare the efficiency. They used a training set of 1000 traffic related tweets and 500 set of data for testing which contained both traffic and non-traffic related tweets. In [4], Sakshi, and Shuchita Upadhyaya, 2016 made data analysis using Apache

Spark using Twitter API. They have retrieved a set of tweets and counted the number of times each word appears in the text file. Then, all the words that are less than a specific number of times in the tweets are filtered out. For the remaining set we count the number of times each letter occurs. The top ten words that are collected during a particular period of time are fetched. The number of a particular “word” being used in the tweets twitted in particular period of time is calculated. Calculate number of a particular “word” being used in the tweets twitted in particular period of time. These steps carry out a defined pattern for the analysis and relevant information from an enormous amount of data from the twitter tweets.

Regarding the classification of tweets, in [5] Sunu Wibirama et al assigned suitable class packaging to every tweet, because related with activity of traffic event or perhaps not. The traffic recognition system or framework was utilized for real-time monitoring of various areas of the street network, taking into account detection of traffic occasions just almost in actual time, regularly before online traffic-news sites. They employed support vector machine like a classification unit and accomplished a great accuracy value of ninety five. 75% by attempting a binary classification issue. Sandeep G Panchal and Prof. R. S. Apare, 2017 in [3], using HDFS (Hadoop dynamic file subsystem) is used for fast processing and storing high amount of data in to HDFS. Web service is taking twitter tweets as input and classifies the traffic related tweets from the twitter. For the classification NLP algorithm is used for classification. The main aim is detection traffic related event from social network. It acts as multi-class classification which is recognizing traffic, non-traffic due to the crash or congestion and traffic due to the external events. The system detect traffic event in real-time. It is the android application in that user has to search location and get route.

Regarding providing interface and notification to users, Kavita Sawant, Shital Pawar et al in [2] created android application where user need to login to the android application and then user can search path and they can see traffic on that way which will be displayed on the map and user can enter the source and destination and obtain an alternate route based on the data obtained through Twitter.

IV. INFERENCES

From going through related works on our project and above mentioned research paper we have derived following inferences. Although work has been done to detect traffic through social network data, but this feature is available to users through Google Maps. Instead, it is important to detect events causing this traffic which involves multi-class classification which can be achieved efficiently through SVM algorithm. No related work is done on Indian cities about traffic event detection. Moreover, in some papers only classification step is completed. No mechanism for notifying the output to users is implemented.

V. PROPOSED SYSTEM

To generate Traffic Event Detection System, we are collecting real-time tweet stream using Twitter API in a text document which will be input to the system. Further, the system is

divided into three modules namely Preprocessing, Classification and Web Application. In preprocessing, first filtration is done which removes unwanted texts such as hashtags, user accounts. Then, we tokenize the filtered text. The tokenized text is further filtered to carry out stop-word filtration as they do not convey any information. In the next step of preprocessing, suffixes and prefixes are removed to fetch root word. This step is called stemming. After, the tweet preprocessing, training and testing of SVM and Logistic Regression algorithm is done to find out their efficiency. Using, the more efficient algorithm, classification is done to find out reasons causing the traffic namely due to accident, landslide or general reasons. The output of the system will be in form of web application. This web application will have Google Maps API embedded into it, from which user can enter source and destination into it. Depending on the location of the fetched tweet, the event is displayed according to the user’s search.

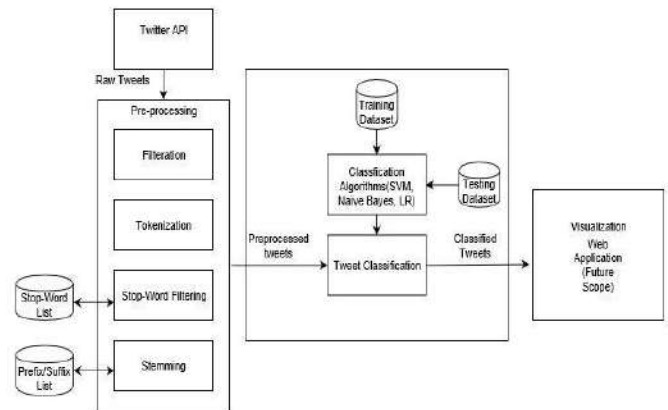


Fig.1 Proposed System architecture

VI. METHODOLOGY

System works on following stages.

1. Extraction of Tweets

Twitter provides Twitter API to extract its data and use it for analysis purpose. To use these API’s, first we need to create a developers account. Following steps are used to create developer’s account

1. Visit the Twitter Developers’ Site dev.twitter.com
2. Sign in with your Twitter Account
3. Go to apps.twitter.com
4. Create a New Application by clicking on the big “Create a new application” button.
5. Fill in your Application Details
6. Create Your Access Token
7. Choose what Access Type You Need
8. Make a note of your OAuth Settings which contains Consumer Key, Consumer Secret Key, OAuth Access Token Key, and OAuth Access Token Secret Key.

1.1 Tweepy API:

Tweepy is easy to use python library to use Twitter API. Tweepy supports accessing Twitter via Basic Authentication and the newer method, OAuth. OAuth only requires users to enter keys mentioned in Step 8 of above process in the code to access the tweets. This prevents writing of passwords in the coding part of this phase. Sample retrieved single tweet is shown below:

2. Preprocessing

A single tweets retrieved from API contains variety of information such as user account name, created_at, id-the tweet identifier, text, screen name, location, comment etc as shown in the figure above A sample tweet retrieved from API looks as shown in above figure. Thus, it is impossible to use this data for data analysis or for applying it on classification algorithms. Hence, the next step after extraction of tweet is preprocessing. .

Pre-processing module is further divided into 4 sub phases:

1. Filtration of Input
2. Token Extraction
3. Stop Word Analysis
4. Stemming

2.1 Tokenization

Tokenization is the process of breaking up a sequence of strings into pieces such as phrases, symbols, words, keywords, and other elements called tokens. Tokens can be individual words, phrases or even whole sentences. In the process of tokenization, some punctuation marks are discarded and the tokens become the input for another process like parsing and text mining. Tokenization relies mostly on simple heuristics in order to separate tokens by following a few steps:

- Tokens or words are separated by whitespace, punctuation marks or line breaks.
- White space or punctuation marks may or may not be included depending on the need.
- All characters within contiguous strings are part of the token. Tokens can be made up of all alpha characters, alphanumeric characters or numeric characters only.

2.2 Filtration of Input

In this phase the input text is completely filtered so that unwanted characters and symbols are removed from the text. Extracted tweets contain hashtags, punctuation mark which we don't require for classification and hence can be filtered out. Moreover, we only need tweet's text, language and location, rest of the things will be removed in this step itself. Algorithm for filtration of input:

1. Accept the single text document as input.
2. Remove punctuation marks and hashtags.
3. Remove all other field expect text body, Lang and location.

After this step, the output will be as follows, as compared to fig2:

```
<created_at: Sat Oct 06 12:25:02 +0000 2016>
<id: 1048549691791736832 id_str:1048549691791736832>
<text: Traffic Vashi bridge collapsing flyover: https://t.co/voOxBLy0Yeg Comments: https://t.co/z30eO1z8D9b>
geo":null, coordinates: null, place: null
<lang:"en", timestamp_ms:1538828702291>
```

Fig.2 Tweet Structure after filtration

2.3 Stop-Word Analysis

Stop-words are the words which acts as fillers in a sentence and does not contain any knowledge and information. Hence, it is better to remove them before carrying out text analysis. Removal of stop-word increases searching performance. However, there is no universal list of stop-word present, but larger the list used as dataset, more accurate stop word analysis will take place.

Algorithm for stop-word analysis is as follows:

Step1: The target document text which is tokenized into individual words.

Step2: Save all the list of stop-word in a separate file called as stop-word dataset.

Step3: Initialize pointers to the start of the file

Step4: While tokenfilepointer! = EOF and StopWordFile Pointer! =EOF

- a. Compare token with stopword
- b. If match is found, then remove that token else Increment the StopwordFilePointer
- c. If no match is found and StopwordFilePointer= Then increment tokenfilepointer and repeat the steps again.

Step5: Resultant text devoid of stopwords is displayed.

Text Field after Stopword analysis will give following output for the given tweet.

```
Traffic at Vashi bridge due to collapsing of flyover  →  Traffic Vashi Bridge collapsing flyover
```

Fig. 3 Text field of Tweet after Stopword Analysis

2.4 Stemming

Stemming is the process of removing suffix and prefix from words to derive the root word from it. A file contains which consist of a list of prefixes and suffix which compares with each word and removes the occurring suffix and prefix. However, the stem is not necessarily the root of the word, for eg. In the context of machine learning based NLP, stemming makes your training data denser. It reduces the size of the dictionary (number of words used in the corpus) two or three-fold. Having the same corpus, but less input dimensions, ML will work better.

Algorithm for stemming:

Step 1: Make corpus of all the possible prefixes and suffixes in a dataset.

Step 2: Read all the valid tokens one by one.

Step 3: While tokenfilepointer AND suffixfilepointer! = EOF.

- a) Find ending of each token and compare it with suffix corpus.

- b) Replace the suffix with NULL character
- c) Update the token which is changed

Step 4: Repeat Step 3 for all tokens.

Step 5: End

After this stage, output from previous stage as shown in Fig. 3 will be as modified as follows:



Fig.4 Text field of Tweet after Stemming

With this stage, preprocessing phase is complete. Now the processed tweets are sent for classification phase. Classified into classes: traffic-related and non-traffic related. Further, traffic-related data will be divided into three-classes: accident, landslide and general-traffic

3. Classification

In this stage, tweets are classified into classes: traffic-related and non-traffic related. Further, traffic-related data will be divided into three-classes: accident, landslide and general-traffic. To do this we have two approaches:

1. Dictionary based: Creating dictionary of all word occurring in traffic related tweets. If number of words from this dictionary occurs more than a threshold value, that tweet is traffic related.

2. Algorithm based: There are machine learning classification algorithm like Naive Bayes, Logistic Regression, Support Vector Machine etc. which can classify data directly.

We used second approach in our paper as Dictionary based method is too tedious. Dictionary may not contain all the words which may impact accuracy.

We use two classification algorithms namely Logistical Regression and Support Vector Machine in our system. Initially both these algorithm is trained using a sample dataset. The sample dataset which is used for training and testing looks as follows

TrafficPrediction	Id	Tweet
1	1.58849570801531053074	As highway construction heats up, please slow down and be aware in #Nork2zones. Work.Zone safety is in your hands!
1	1.5902255612117708345	Crash 16 Telegraph near Jay Road in Redford Township affecting Left Lane. Traffic slow from Plymouth Road. #wvttraffic
1	1.5886762813195794179	Incident on #Line 6 both Dir/Both Dir at Neptune Avenue Station
1	1.5900826637440447937	Updated: Crash in Palm Beach on highway north at Exit Yamato Rd, left lanes blocked. Last updated at: 10:45PM.
0	0.5840215804671057921	Shared spaces and new amenities are redefining the way offices facilitate community and creativity. #e!

Fig. 5 Sample dataset

4. Web Application

This is planned as a future scope of our project. Once, the events are detected, they need to be displayed on an interface. For this, we will create a web-application which will have Google Maps API embedded in it. Through Map, we will be able to detect traffic automatically and display the retrieved tweets on the maps. Thus, user will be able to get traffic information from Maps API and events due to which traffic is

caused will be displayed by our analysis. User need to provide source and destination address and search the route to get the information. Otherwise, if GPS in device is enabled, user will get notification automatically. For this, however, user needs to provide required device permissions.

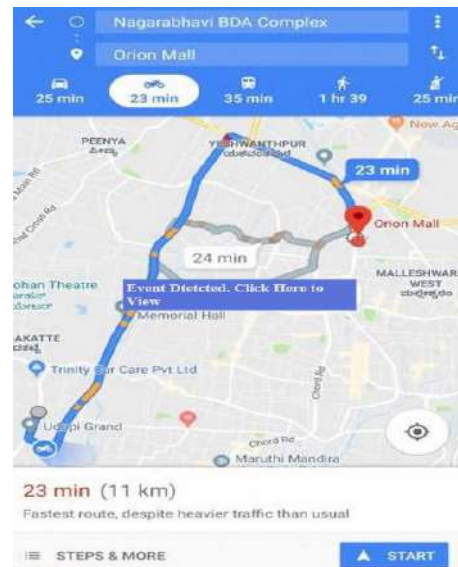


Fig.6 Event Detection

VIII. RESULT/ANALYSIS

The dataset will be divided into training and testing set in the ratio 80:20, and will be used for training the testing two algorithms namely Logistical Regression and Support Vector Machine. The algorithm with higher accuracy and efficiency will be used in our system. Thus, we will have a definite proof about the algorithm, which we are using the system is most accurate

IX. APPLICATION

Data extracted from social networking sites have wide variety of application ranging from data analysis for elections predictions, movie reviews etc. Moreover, Natural Language Processing has famous application for bot to communicate in human language. Machine Learning classification algorithm predicts classes for data analysis. Following are the application of the system:

1. Tweet analysis have wide variety of application to predict movie reviews sentiment, target interested customers, even predicting election results by analyzing user's opinion.
2. Traffic event detection will give insights to users of how long the traffic will remain. In Google Maps, although traffic is detected but there is no information as to why traffic is present at that location. If the traffic is due to serious event such that collapsing a flyover, then traffic at that location will take long time to disperse. Hence, user can altogether avoid going to that road. Thus, detailed traffic information can be obtained though this system.

3. This system acts a skeleton on which we more work can be done covering larger area and events. Thus, this system can be applied on a larger scale.

X. SUMMARY

The proposed system has presented to detect traffic events using Twitter data analysis. The system accepts input a text file which contains tweets retrieved in unstructured format and transforms into structured format using preprocessing. Using classification algorithm, the system classifies the structured tweet into traffic-related and non traffic related. The traffic related tweet is further classified into events such as accident, landslide and general events. This work can be used as an update or feature enhancement for Google Maps API because currently it does not report traffic events to its users.

XI. ACKNOWLEDGEMENT

It is a great pleasure and moment of immense satisfaction for us to express my profound gratitude to our Project Guide, Prof. Sagar Kulkarni whose constant encouragement kept us motivate enabled us to work enthusiastically. We are thankful to Dr. Sharvari Govilkar, H.O.D, Information Technology Department and Dr. Sandeep Joshi, Principal, Pillai College of Engineering, New Panvel, for their for providing an outstanding academic environment and platform and adequate facilities. We are thankful to all our teachers who are willing to always help us whenever needed.

X. REFERENCES

- [1] Medha A. Shah, Ketan R. Pandhare, 2017, "Real time Road Traffic Detection and Apache Spark", International Conference on Inventive Communication and Computational Technologies, ICCICCT 978-1-5090-5297-4/17, pg 445-449.
- [2] Mrs. Kavita Sawant, Miss. Shital Pawar, Miss. Poonam Jadhav, Mrs. Sayali Vidhate, Mrs. Nirasha Bule, Mrs. Snehal Patil, May 2017,"Traffic Detection from Real Time Twitter Stream Analysis and Navigation System", IJSEC Vol.7 Issue No.5, Pg.12252-12255.
- [3] Sandeep G Panchal, Prof. R. S. Apare ,2017, "Real Time Traffic Detection using Twitter Tweets Analysis", International Journal of Engineering Trends and Technology (IJETT) – Volume 47 Number 8 May 2017, Pg.458-462.
- [4] Sakshi, Shuchita Upadhyaya, December-2016 , "Twitter Streaming API Using Apache Spark in Big Data Analytics", International Journal of Scientific & Engineering Research, Volume 7, Issue 12, 354ISSN 2229-5518, Pg. 354-359
- [5] DWI Aji Kurniawan1, Sunu Wibirama, Noor Akhmad Setiawan, 2016, "Real-time Traffic Classification with Twitter Data Mining", 2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE),
- [6] Sweety Kumari1, Firdos Khan, Sheikh Sultan, Ruchita Khandge, Real-Time Detection of Traffic From Twitter Stream Analysis, International Research Journal of

Engineering and Technology (IRJET), Volume: 03 Issue: 04 - 2016, e-ISSN: 2395 -0056, p-ISSN: 2395-0072, pg 2350-2354.

[7] Eleonora D'Andrea, Pietro Ducange, Beatrice Lazzerini, Francesco Marcelloni, "Real-Time Detection of Traffic from Twitter Stream Analysis", IEEE

XII. AUTHORS PROFILE



Mohit A Rai is pursuing Bachelor of Engineering in Information Technology from Pillai College of Engineering, New Panvel. He is currently in fourth year of the course. His areas of interest lie in

Natural Language Processing, Machine Learning, Text Mining, Android Programming and Web development.



Riya Sawant is pursuing Bachelor of Engineering in Information Technology from Pillai College of Engineering, New Panvel. She is currently in fourth year of the course. Her interest lies in text mining,

machine learning and web development.



Sumod Menon is pursuing Bachelor of Engineering in Information Technology from Pillai College of Engineering, New Panvel. He is currently in fourth year of the course.



Devbrat Singh is pursuing Bachelor of Engineering in Information Technology from Pillai College of Engineering, New Panvel. He is currently in fourth year of the course.

Chatbot Based Question Answering System

Giridhar Srinivasan, Voval Jain, Prashant Niladhe, Vishal Gupta, Deven kanse
Students
Pillai College Of Engineering, New Panvel, sector-16

Project Mentor
Prof. Shubhangi Chavan

Abstract— Chatbot presents a new way for individuals to interact with the computer systems. The chatbot provides a chat interface which allows user to ask questions in same way as they would address a human. The System makes use of ontology and NLP to provide the best response by interpreting the input text. Using information retrieval and NLP techniques chatbot identifies question and mines the suitable answer. The implementation of project is to be divided in these phases: Accessing Natural language Query where the input query is read, preprocessed and gets tokenized, next step is feature extraction phase wherein required phrase from query is to be analysed, Finally answer retrieval is carried out according to the query and displayed in the chatbot GUI.

Keywords: Tokenization, Chunker, Ontology, Classifier, Onto triple, Natural language processing, Onto matching,

I. INTRODUCTION

We're in the middle of a chatbot revolution and it won't be long until businesses and brands begin to engage in a battle of the bots. In this next era of consumer engagement, sales and customer service, the winners will be those companies that begin their chatbot journey today not tomorrow

For businesses, it has become necessary to solve the queries and problems of the customers to ensure consumer loyalty along with the brand establishment. And just like the earlier times, man has looked to take help of machines to remove the constraints of human limitations. This time it is the customer service industry which has been revolutionized, and the innovation responsible for this is chatbot. Chatbots are considered the future of customer service and management.

Computer-based chat system is one of the most popular communication methods used in the modern world. As such, there are so many chat-systems available world-wide. These chat systems can be broadly classified into two categories, namely, human-human dialog system and human-computer dialog systems. Both systems enable communication using natural languages such as English. The latter systems generally named as Chatbot. The basic idea of this project came from after the referral of various research papers and understanding the technology used in past project research. We have studied various research papers to understand the techniques of NLP, Tokenization, POS tagging and Ontology. In this chapter relevant technique is literature is reviewed. It describes various techniques used in the work. Identify the techniques that have been developed and present the various

advantages and limitation of these methods are covered in this chapter.

II. SCOPE OF THE PROJECT

Chatbot Based Question Answering System will provide appropriate response to the user given Query by understanding the behaviour of user. The system will collect user information while having interaction with system to provide most accurate response. The response will be fetched from Chatbot API is unstructured and hence need to be processed using various preprocessing techniques. Therefore Answer generation technique is used to generate response. The system will focus on what is stored in ontology to extract the answer. Rest of the Query related task like classification, tokenization, chunking, onto-tripple generation. The system currently focuses on the Hindi language. The system can be integrable with E-commerce sites which will help customer to buy the needed product..

III. RELATED WORK

Large number of research paper and information related to Chatbot, Ontology, preprocessing, and Classification algorithm are available on Web.

To process question provided in Hindi language and retrieve answers for those question, Sharma, Lokesh Kumar, and Namita Mittal[1] have used Named Entity based n-gram approach for their question answering system. For retrieval of answers first question classified and analyzed to generate a proper query. Question classification helps to identify relevant type of answers. Then by using similarity metric relevant document is retrieved which probably contains the answer and at last by using the bigram and NER relevant answers are retrieved for the given question. Overall higher accuracy was obtained by using the bigram approach but accuracy dropped in scenario where synonyms present in document where not matched due to the use of syntactical approach.

A dialogue based question answering system which provides answers related to railway domain in Telugu language is proposed by R. Reddy, N. Reddy and S. Bandyopadhyay [2]. Question answering process is based on keyword approach where input query are tokenized and keyword are extracted using knowledge base related to railways. Tokens generally consist of train names, station names whereas keywords specify when, in, out, go and others present in the query text.

Query frame is extracted by matching it with predefined procedures to generate relevant SQL query. Dialog manager task is to interact with users if more information is needed to execute SQL query to fetch relevant answer to user question.

Wang, Chong, et al [11] has created a Portable natural language interface to Ontologies, name as PANTO which accepts generic natural language queries and outputs SPARQL queries. Based on a special consideration on nominal phrases, it adopts a triple-based data model to interpret the parse trees output through parser. They have used Stanford Parser and multiple existing techniques and tools are integrated to interpret parse trees of natural language queries into SPARQL. To understand sense of the words in the NL queries and WordNet and string metrics algorithms are also integrated.

Architecture for ontology based natural language question is proposed by Raj, P. C. [4] where concept of semantics and ontology is used to facilitate better query construction and extraction of answer. Architecture consists of question processing, document extraction and processing and finally answers processing. Here in the question processing module the question is analyzed using NLP techniques like POS tagger, Parser, NER. In second module relevant documents are retrieved from repository based on conceptual indexing and processed to extract candidate answer set. In answer processing module candidate answers are filtered and finally answer are generated. The literature review shows that most of the existing QA systems are available for English language and some researchers have worked on Hindi, Telgu and Punjabi as Indian regional languages. Most of these algorithms have used Cross Lingular based approach to extract the information. The QA system for Telgu is based on dialogue manger which uses SQL query generator to fetch answer. Most of the existing system mostly provide answers for “what, where, when and who” type of questions only.

Question answering system to produce answer of question in Punjabi and English is proposed by V. Gupta [6]. The system accept query in English or Punjabi language of which stop word is eliminated initially. Then from the query string key terms like noun, adjective, verbs or adverb are extracted. Using dictionary of Punjabi and English language synonyms of key terms is extracted. Finally query is reformulated using the extracted keywords and its synonyms. By using reformulated query various matching web pages are retrieved using a search engine. Extracted documents are summarized based on proximity of key term found in documents and finally candidate answer is provided as per its rank.

IV. INFERENCES

From going through related works on our project and above mentioned research paper we have derived following inferences. Although work has been done to create chatbot system which gives appropriate response to the user, but this system is not available for pure hindi language . Till now the

accuracy for pure hindi language system is less . Moreover, in some papers only classification step is completed. No techniques for generating more accurate answer is implemented.

V. PROPOSED SYSTEM

The system is divided into six modules namely Preprocessing, Pos Tagging, chunker, Ontology, Answer Generation and User Interface. In preprocessing, first filtration is done which removes unwanted texts such as hashtags. Then, we tokenize the filtered text. The tokenized text is further filtered to carry out stop-word filtration as they do not convey any information. In the next step of preprocessing, suffixes and prefixes are removed to fetch root word. This step is called stemming, followed by chunker process which usually selects a subset of the tokens ,Further the process of chunking is carried out to extract noun and verb group from the POS tagged question. Chunked groups can be in the form of proper noun, common noun and verb groups. ontology module look for the appropriate answer in the stored data using onto-tripple since Every question of Hindi language may at least content a subject in it or it can contain combination of subject object and predicate. Subject object and predicated thus contribute for generation of query triples in the question. Query triple thus generated are transformed to onto triple. and then finally the more appropriate information is extracted, this information is used by Answer Generation Module to generate the quick and easy response, After processing the data the machine gives an output which is displayed on the user interface.

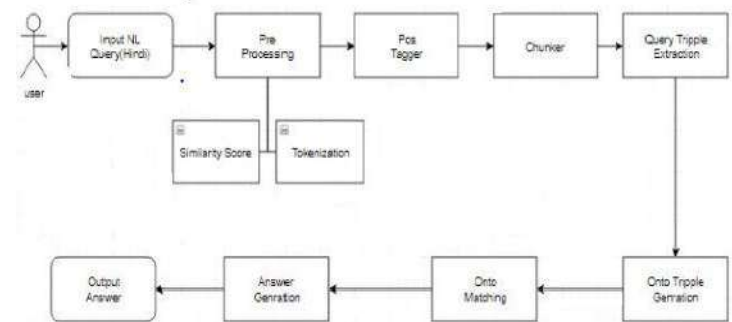


Fig 1.ChatBot Based Question Answering System

VI. Methodology

System works on following stages.

1. Preprocessing

- Preprocessing consist of two modules,
 1.Similarity Score
 2.Tokenization

Tokenization

Tokenization is the process of breaking up a sequence of strings into pieces such as phrases, symbols, words, keywords, and other elements called tokens. Tokens can be individual words, phrases or even whole sentences. In the process of tokenization, some punctuation marks are discarded and the tokens become the input for another process like parsing and text mining. Tokenization relies mostly on simple heuristics in order to separate tokens by following a few steps:

- Tokens or words are separated by whitespace, punctuation marks or line breaks.
- White space or punctuation marks may or may not be included depending on the need.
- All characters within contiguous strings are part of the token. Tokens can be made up of all alpha characters, alphanumeric characters or numeric characters only.

Example: एलेग्रिया में सांस्कृतिक कार्यक्रम क्या हैं ?

Tokenization:

1. एलेग्रिया
2. में
3. सांस्कृतिक
4. कार्यक्रम
5. क्या
6. हैं

2. POS Tagger: Once tokens are generated we apply POS [Part-of-Speech] tagging technique on each token using our POS tag database. POS tag DB will contain tags, from which our system will extract an equivalent tag defining the token and assign the tag to the token.

POS Tagging Output: 1. एलेग्रिया - NNP 2. में - PSP 3. सांस्कृतिक - NN 4. कार्यक्रम - NN 5. क्या - PIN 6. हैं - NNP

Stemming

Stemming is the process of removing suffix and prefix from words to derive the root word from it. A file contains which consists a list of prefixes and suffix which compares with each word and removes the occurring suffix and prefix. However, the stem is not necessarily the root of the word. For eg. In the context of machine learning based NLP, stemming makes your training data denser. It reduces the size of the dictionary (number of words used in the corpus) two or three-fold. Having the same corpus, but less input dimensions, ML will work better.

Algorithm for stemming:

Step 1: Make corpus of all the possible prefixes and suffixes in a dataset.

Step 2: Read all the valid tokens one by one.

- Step 3:** While tokenfilepointer AND suffixfilepointer != EOF.
- a) Find ending of each token and compare it with suffix corpus.
 - b) Replace the suffix with NULL character
 - c) Update the token which is changed

Step 4: Repeat Step 3 for all tokens.

Step 5: End

Output-1. एलेग्रिया - NNP 2. में - PSP 3. सांस्कृतिक - NN 4. कार्यक्रम - NN 5. क्या - PIN 6. हैं - NNP

3. Chunker & Query Triplet

Chunker contains Subject , object and predicate called as There are 13 types of chunk tags.

Output-Noun Group 1: एलेग्रिया NNP, हैं NNP

Noun Group 2: सांस्कृतिक कार्यक्रम NN

4. Ontology

Determines properties of object and relation between object. Properties of object is derived. Concepts and properties are classes in the ontology. Data properties are relation between classes and data values where domain is a class and range is set of values.

In the following sample, the root word here is alegria.

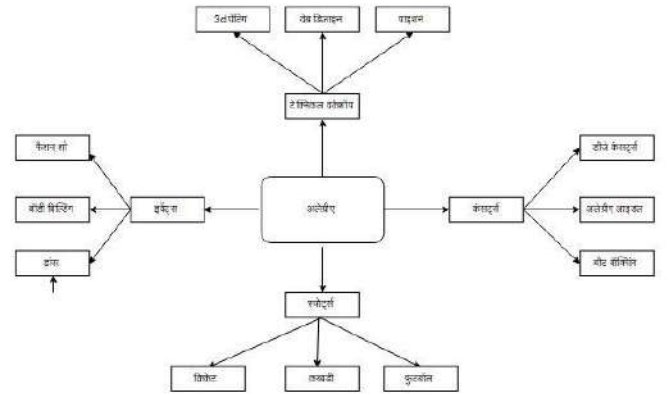


Fig2. Ontology Structure

Onto triplet:

Matching of onto terms of words with those stored in ontology is done which leads to retrieval of accurate answer for the given sentences.

Example- क्या(एलेग्रिया,सांस्कृतिक कार्यक्रम)

In the above example the words like एलेग्रिया, सांस्कृतिक,

कार्यक्रम are matched with the words stored in ontology.

VII. Result Analysis

The extracted answer for given question is

Output- एलेग्रिया में सांस्कृतिक कार्यक्रम टेक्निकल वर्कशॉप,कंटेंट्स तथा सांस्कृतिक प्रसंग हैं।

VIII. APPLICATION

There are various applications of this chatbot system.

The application are as follow.

- They can carry out a lot of tasks depending on how you design it. For example, any task from ordering a pizza, booking a hotel room to telling jokes on demand .
- A Question answering (QA) system is a system that automatically answers questions posed by humans in a natural language, utilizing natural language processing (NLP), and a knowledge base (a structured database of information of which the QAS can query).
- Chatbots are totally different from a question answering system. A chatbot not only give answers to the questions of the customers but it also perform various other works like generating leads for businesses,increasing sales , look after customer queries ,services & satisfaction, promoting products etc.
- Since Facebook Messenger, WhatsApp, Kik, Slack, and a growing number of bot-creation platforms came online, chatbot is used on a large scale in our daily life.

IX. CONCLUSION

The proposed system has presented to make a conversation between both human and machine. It takes input as a question from user, then accordingly machine gives appropriate answer in response. The Login Module will allow the user into the system and provide to enter the input question in hindi language. Then pre-processing module classifies and filters the input sentence followed by chunker process, ontology module look for the appropriate answer in the stored data using onto-tripple and then finally the more appropriate information is extracted, this information is used by Answer Generation Module to generate the quick and easy response, After processing the data the machine gives an output which is displayed on the user interface. Overall this system provides accurate reply to the user.

X. ACKNOWLEDGEMENT

It is a great pleasure and moment of immense satisfaction for us to express my profound gratitude to our Project Guide, Prof. Shubhangi Chavan whose constant encouragement kept

us motivate enabled us to work enthusiastically. We are thankful to Prof. Sharvari Govilkar, H.O.D, Information Technology Department and Dr. Sandeep Joshi, Principal, Pillai College of Engineering, New Panvel, for their for providing an outstanding academic environment and platform and adequate facilities. We are thankful to all our teachers who are willing to always help us whenever needed.

XI. REFERENCES

- [1] Sagar Pawar, Omkar Rane, Ojas Wankhade, Pradnya Mehta. "A Web Based College Enquiry Chatbot with Results." International Journal of Innovative Research in Science, Engineering and Technology Vol: 7, Issue 4, April 2018.
- [2] Sharvari S. Govilkar and J.W. Bakal. "QUESTION ANSWERING SYSTEM USING ONTOLOGY IN MARATHI LANGUAGE." International Journal of Artificial Intelligence and Applications (IJAIA), Vol.8, No.4, July 2017.
- [3] Vibhor Sharma, Monika Goyal,Drishiti Malik. "An Intelligent Behaviour Shown by Chatbot System" International Journal of New Technology and Research, vol-3 Issue-4, April 2017.
- [4]Garima Nanda, Mohit Dua and Krishma singla. "Hindi Question Answering System using Machine Learning Approach." Department of Computer Science and Engineering, Banasthali Vidyapith, Jaipur, Rajasthan, 2016.
- [5]Bayu Setiaji, Ferry Wahyu Wibowo. "Chatbot Using A Knowledge in Database.Human-to-Machine Conversation Modeling." Department of Informatics Engineering STMIK AMIKOM Yogyakarta Yogyakarta, Indonesia, 2016.
- [6]Seena I T , Sini G M, Binu R. "Malayalam question answering system." M Tech Computational Linguistics, Dept. of Computer Science and Engg. Govt. Engineering College, Sreekrishnapuram, Kerala, 2015.

Abstract:

Security has been playing a key role in many of our places like home, offices, institution, suitcases, etc. In order to avoid intrusion from unauthorized person into these places a portable smart lock is proposed. Biometric systems and facial recognition have overtime served as robust security mechanisms in various domains. Fingerprint is most widely used form of biometric identification. Project builds an IOT based portable smart lock which can be opened through various means such as biometric fingerprint and facial recognition via mobile application using Wi-Fi or Bluetooth module. Database will be used to store the records of authorized person to unlock the lock. When an unauthorized person tries to unlock the lock a push message will be send to the owner of the lock and subsequently log of the same will be saved in the database .This database will be stored on web-server. Hence the lock will be unique combination of various aformentioned security features providing solution to problem of security.

Keywords:

Facial recognition, fingerprint,bluetooth,portable,database,security

*Corresponding Author Email:manthanparvadia@gmail.com / onkar22pokharkar@gmail.com / ayushshetty17@gmail.com / shindeshubh.ss@gmail.com

Phone: 9920722291 / 9987292363

I. INTRODUCTION

In this modern world crime has become ultra modern too! In this current time a lot of incident occurs like robbery, stealing unwanted entrance happens abruptly. So the security does matters in this daily life. People always remain busy in their day to day work also wants to ensure their safety of their beloved things. Sometimes they forget to look after their necessary things like keys, wallet, credit cards etc[1].

The technology of keys and locks remained the same for the last century while everything else is evolving exponentially. So why not use current technologies and apply it with old ones to build something new and innovative[2].

Recently, the Internet was enhanced, and everything was connected to it (phones, televisions, laptops, tablets, cars and so on...). This was done because we wanted to make systems “smarter”, in other term “more productive”. Why not do the same thing with Locks? Enhancing the locks mechanism by connecting them to the internet, making them more robust and productive.Today, the number of mobile device users including smartphone users has rapidly been increasing worldwide, and various convenient and useful smartphone applications have been developed. Now smartphones are not only used to send and receive phone calls, send text messages, and perform mobile banking operations, but they also are used to control various other devices in our real everyday lives. Through a mobile operating system and internal applications, we can remotely control a

variety of external devices such as TVs,projectors, computers, cars, etc[2].

Biometrics are automated methods of recognizing a person based on a physiological or behavioral characteristic. Among the features measured are; face, fingerprint, hand geometry, iris, retinal, signature, and voice. Biometric technologies are becoming the foundation of an extensive array of highly secure identification and personal verification solutions. As the level of security breaches and transaction fraud increases, the need for highly secure identification and personal verification technologies is becoming apparent.

In this paper, a new system is designed which would be a combination of two biometric factors (face and fingerprint) which would be integrated in a single system.The user can unlock the lock either through fingerprint present on the lock or face detection via mobile application. The system would be integrated in such a way that the lock can be carried any time anywhere thus increasing its application areas and making it portable.

II. METHODOLOGY

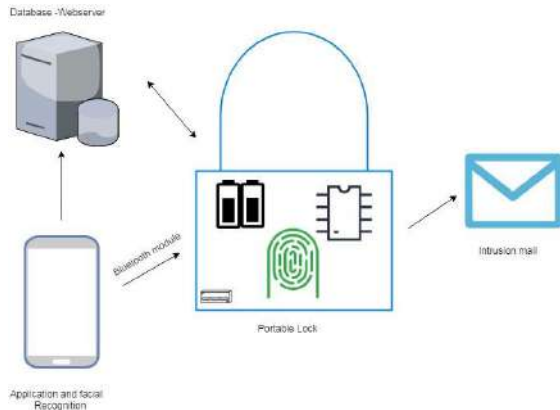


Fig. 1 System Architecture

A. System Architecture

i) Server:

The server provides two things:

- a) Database: It stores the log of entry and intrusion detection.
- b) Web server: It manages the database and communicates with other components request/response.

ii) Mobile Application:

Android mobile application is developed to allow users to register to use the system and access the features of the system. The owner can authenticate other users to use the system and its features. Android library is used for face recognition. The application communicates with the lock via bluetooth to unlock the lock using signals. The application can access logs too.

iii) The lock:

The lock is portable and can be carried anywhere anytime. The lock consist of following components in it:

- a) Fingerprint scanner: The fingerprint scanner scans the print of the user placed on the fingerprint scanner with the fingerprints stored in the database.
- b) Arduino uno: It is used as micro controller. Controls other components by sending control signals. Controls bluetooth and wifi capabilities.
- c) Battery: 9V-12V batteries are used to provides power supply to the fingerprint scanner.

- d) Usb port: usb port is used for charging the batteries.

iv) Email/Msg:

When an unauthorized person tries to unlock the lock using fingerprint or facial recognition a email/msg is send to the owner and log is maintained of the same.

B. Features

i) Multiway unlocking system:

The system can be unlocked either by facial recognition or fingerprint whichever is convenient for the user at that moment.

ii) Intrusion detection system:

The system sends an email/msg to the owner if the lock is tried to be unlocked by unauthorized user.

iii) Logs:

The system keeps recordings of the log by maintaining the history of lock/unlock operations.

iv) Availability:

Android application features can be availed and accessed anywhere anytime and authenticate other users to access the lock.

C. System Methodology

i) Registration:

User registers himself using the android mobile application. Logins himself and registers face image and fingerprint which are to be recognized as authentic. The owner can register other users as well and store face images and fingerprint which are to be recognized as authentic.

ii) Operation:

User unlocks the lock either using facial recognition or fingerprint scanner. If the user is authorized the lock unlocks otherwise after predetermined attempts intrusion mail/msg is send to the owner.

III. EXPERIMENTATION

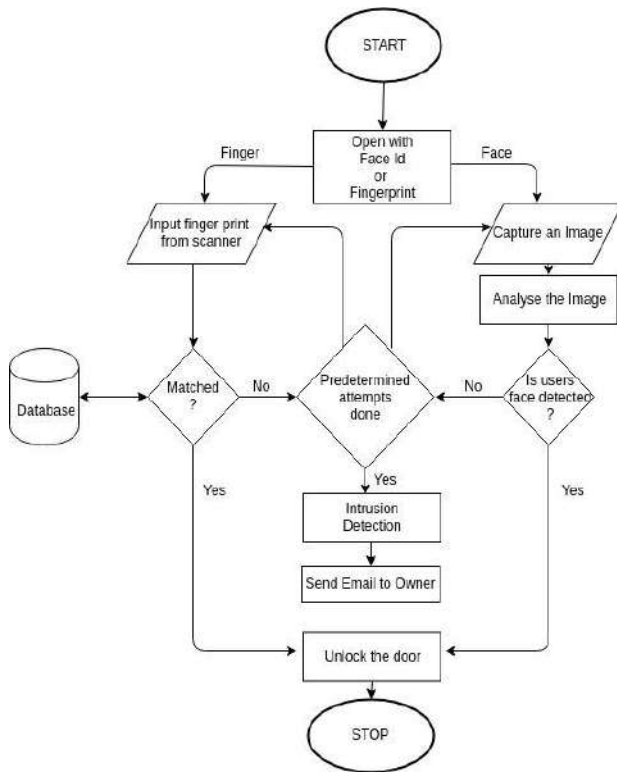


Fig. 2 System working flowchart

Step 1: Start

METHOD 1:

- Step 2: User // who will try to enter biometric details
- Step 3: Finger Print //user will put the finger on fingerprint scanner
- Step 4: Fingerprint scanning // System will match the input with existing fingerprint in the database
- Step 5: if match found the lock is unlocked
- Step 6: Else go to step 8.
- Step 7: Entry in register // users check in time is entered in register.

METHOD 2:

- Step 2: User // who will try to enter his details
- Step 3: Face Id //user will scan his face on the camera.
- Step 4: Face Recognition // System will try to recognise the authentic person
- Step 5: if match found the lock is unlocked
- Step 6: Else go to step 8.
- Step 7: Entry in register // users check in time is entered in register.

- Step 8: If any of the step 3 of method 1 or 2 has been attempted for five times unsuccessfully then go to Step 9
- Step 9: Intrusion will be detected.
- Step 10: Email/message will be forwarded to owner.

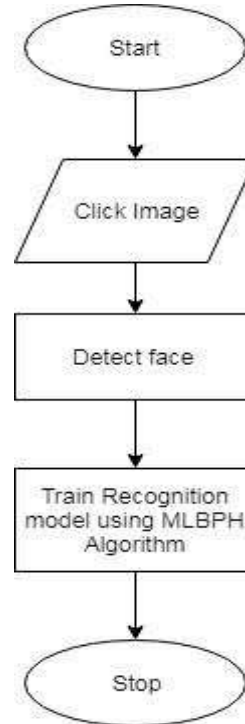


Fig. 3 Facial recognition model training flowchart

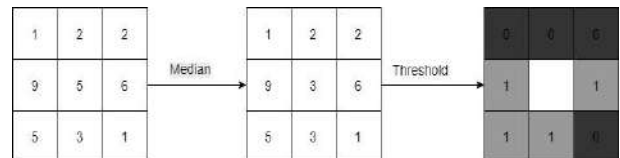


Fig. 4 MLBP operator[3]

Step 1: Capture face image.

Step 2: Use Haar Cascades Classifier with AdaBoost algorithm for Face Detection.

Step 3: If face is detected at Step 2 proceed to Step 4 else terminate.

Step 4: Divide the face image into several blocks.

Step 5: At each block calculate median of all gray values and replace the center value with the median.

Step 6: Consider the center value as threshold of window and compare all other values with it.

Step 7: Calculate Histogram for each block.

Step 8: Concatenate the entire block MLPBH.

Step 9: Compare the MLBPH of current image with MLBPH of saved image.

Step 10: If match found goto Step 11 else terminate.

Step 11: Face recognized successfully.

IV.RESULT AND DISCUSSION

A database of 200 different people with 6 images of each person would be created. Each person's different characteristic images would be selected as test images for training. The experiment will be performed ten times and average of the experiment will be noted.

Table 1 Expected result after training

Characteristics	Correct times	Wrong times	Recognition rate
Illumination change	1167	33	97.25%
Attitude change	1186	14	98.83%
Face proportion change	1111	89	92.60%

The above table shows expected recognition rate when trained with MLBPH algorithm when following characteristics are taken into consideration

V.CONCLUSION

The main advantages of using this system are:





- A. The lock is portable can be carried anywhere.
- B. No issue of power failure since battery is used.
- C. No manual errors.
- D. MLBPH algorithm used for face recognition overcomes the recognition rate disadvantages of LBPH.
- E. Combination of fingerprint authentication and facial recognition overcomes each others disadvantages providing absolute solution to problem of security.


The solution proposed in this paper is a combination of two biometric factors (facial recognition and fingerprint) in both the system overcomes the disadvantages of each other. All notification and data updates across the system are real time since the components of the system are synchronized. The system would be integrated in such a way that the lock can be carried any time anywhere thus increasing its application areas and making it portable. Hence the system is effective yet simple to use solution for security.

VI.REFERENCES

- 1.Md. Nasimuzzaman Chowdhury, Md. Shiblee Nooman², Srijon Sarker³, "Access Control of Door and Home Security by raspberry pi through internet," *International Journal of Scientific & Engineering Research*, Volume 4, Issue 1, 2013.
- 2.Abdallah Kassem and Sami El Murr Georges Jamous, Elie Saad and Marybelle Geagea, "A Smart Lock System using Wi-Fi Security," 2016 3rd International Conference on Advances in Computational Tools for Engineering Applications (ACTEA).
- 3.XueMei Zhao, ChengBing Wei*, "A Real-time Face Recognition System Based on the Improved LBPH Algorithm," 2017 IEEE 2nd International Conference on Signal and Image Processing.
- 4.Varad Pandit, Prathamesh Majgaonkar, Pratik Meher, Shashank Sapaliga, Prof.Sachin Bojewar, "Intelligent Security Lock," International Conference on Trends in Electronics and Informatics ICEI 2017.
- 5.A. Aditya Shankar¹, P.R.K.Sastry, A. L.Vishnu Ram³, A.Vamsidhar⁴, "Finger Print Based Door Locking System," *International Journal Of Engineering And Computer Science* ISSN:2319-7242 Volume 4 Issue 3 March 2015.
- 6.G. Sowmya¹, G. Divya Jyothi¹, N Shirisha¹, K Navya¹, B Padmaja², "Iot Based Smart Door Lock System," *International Journal of Engineering & Technology*, 7 (3.6) (2018) 223-225.
- 7.Bhalekar Pandurang¹, Jamgaonkar Dhanesh² Prof. Mrs. Shailaja Pede³, Ghangale Akshay⁴ Garge Rahul⁵, "smart lock: a locking system using bluetooth technology & camera verification," *International Journal of Technical Research and Applications* e-ISSN: 2320-8163, www.ijtra.com Volume 4, Issue 1 (January-February, 2016), PP. 136-139.
- 8.Mrutyunjaya Sahani, Chiranjiv Nanda, Abhijeet Kumar Sahu and Biswajeet Pattnaik, "Web-Based Online Embedded Door Access Control and Home Security System Based on Face Recognition," 2015 International Conference on Circuit, Power and Computing Technologies [ICCPCT].

Author Biographical Statements

	<p>Payel Gaurav Thakur is a assistant professor of computer engineering department in pillai's college of engineering (panvel).She has completed her ME & Be from pillai's college of engineering (panvel),University of Mumbai</p>
	<p>Ayush Shetty is a final year undergraduate student, Department of Information technology from pillai's college of engineering (panvel),University of Mumbai.He is a member of computer society of india(CSI).He was appreciated for his work on the project Android application for Swachh Bharat Mission(Third Year).</p>
	<p>Manthan Parvadia is a final year undergraduate student, Department of Information technology from pillai's college of engineering panvel,University of Mumbai.He was appreciated for his work on the project Suspicious Email Detection(Third Year).</p>
	<p>Onkar Pokarkhar is a final year undergraduate student, Department of Information technology from pillai's college of engineering (panvel),University of Mumbai.He was appreciated for his work on the project Android application for Swachh Bharat Mission(Third Year).</p>

	<p>Shubham Shinde is a final year undergraduate student, Department of Information technology from pillai's college of engineering (panvel),University of Mumbai.He is a member of computer society of india(CSI).He was appreciated for his work on the project Android application for Swachh Bharat Mission(Third Year).</p>
--	---