Journal of
Information Technology

Volume 7, Issue 1, 2019-20

JIT

Volume 7

Issue 1

2019-20

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 7, Issue 1 2019-20

Editorial

Dr. Sandeep M. Joshi
Principal, PCE

I am very pleased to note that Volume 8 Issue 1 brought out is appreciated very well. Hearty congratulations to the editorial team.

The role of the teachers is to nurture the skills and talents of the students as a facilitator for research and development. The JIT from the department of Information Technology is going to showcase the strength of our Institute.

I extend best wishes for the success of this endeavor.

Editorial

Dr. Satishkumar L Varma
Editor-in -Chief

Dear faculty and students of Pillai College of Engineerig,

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 Natural language Processing, Machine Learning, IoT, Security and Data Mining. This journal discusses at length the application of natural language processing to various problems such as disease prediction for fruits, plagiarism detection, syntax detection for regional languages, safe road for driving and object detection. It also contains papers demonstrating the use of IoT in the applications such as smart healthcare, traffic signs and signal detection.

This issue covers fourteen 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 Natural Language Processing, IoT and Data Mining.

Thankful to all our studetns and teachers for putting extra efforts to pring this issue.

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

Contents

Contents

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 31 refereed Journal papers and more than 35 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 has received her Ph.D in Computer Engineering with research area Usability Engineering in Agriculture Domain in 2019. 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 22 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 Engineeing 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 purblications and has attended 6 FDP. Her area of interst includes Operating system, Cloud Computing, Big Data Analytics and Distributed Systems.

Madhu Nashipudimath has received her Ph.D degree from P.A.H.E.R University, Udaipur in the field of big data analytics in 2020 and received 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 35 papers in International and National Journals and Conferences and also attended more than 38 conferences and workshops. She has participated as organiser, speaker and participant in 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 has been a reviewer for several peer reviewed conferences and publications in International and National conferences . She has reviewed research papers of reputed journals like International Journal Big Data and Springer. Her fields of interest are Data Mining, Big data, Information Retrieval and Fuzzy logic.

HIRE EASY- IDENTIFY YOUR WORKPLACE

Dr. Sushopti Gawade[1] ,Sonal Kawle[2], Tushar Hadawale[2], Rutuja Sawant[2]
[1]Professor, [2]Students.
[1,2]Department of Information Technology,
[1,2]Pillai College Of Engineering (PCE), New Panvel- 410206
[1]sgawade@mes.ac.in, [2]kawlesh15it@student.mes.ac.in, [2]tusharh78@student.mes.ac.in,
[2]rutujashsaw16de@student.mes.ac.in

Abstract: As new trends emerge in the field of Information Technology, a vast number of job opportunities are getting created everyday. A jobseeker, especially a fresher may get confused as to what company is he/she best suited for. In this scenario, there comes a need for a solution which can predict using the power of machine learning, which company the jobseeker is most suited for. To find a solution for this problem, the HireEasy app is created. The app takes into account a jobseeker's skill-set, the placement data of various companies then using an algorithm predicts the chances of that candidate getting selected in those companies. As the tagline suggests, the application helps the candidate identify what company is suitable for him/her.
In the future, the application will get better at predicting (supervised learning) the results because of the growing dataset. The machine learning algorithm used for the prediction of results is Support Vector Machine algorithm. It takes into account the training dataset and outputs an optimal hyperplane which categorises new examples.

Keywords: Machine learning, Prediction system, Supervised learning, Support Vector Machine algorithm.

1. Introduction: In the traditional job search and recruitment process, the only way to select the best qualified candidate is to have a pool of eligible applicants, made possible by drawing the interest of individuals in the market. An employer seeking a candidate for a particular job post creates an advertisement, stating his/her requirements (qualification, experience, skillset etc) and then posts that ad in either newspapers or on online job portals. Now, there can arise possible problems that can hinder this recruitment process.

Sometimes, the employer does not clearly mention all the necessary details or may leave any essential skills unmentioned. On the other hand, there are many candidates who apply to job postings for which they are not suited for. There could be many reasons for this, which could waste time and money for both, employer as well as the candidate. Machine learning (ML) gives us the power to derive meaningful or useful data from a

pool of available data. Based on the given data by the candidate, provided that it accurate as per him/her, and the placement records of the companies in the past , the machine learning algorithm generates a personalised result for the candidate, which the candidate can refer to before applying to any company. Also, the employer can look at the prediction results of the various candidates and select a candidate having a high predictability rate, which can help the employer to select the best candidates from the remaining candidates for the job. As the dataset increases, the application will be better able to predict the results of job suitability.

The application can help to bridge the gap between the employers not being able to find the right candidates for the job and the candidates applying for positions not suited for them.

2. Literature review:

A. Job procurement: Old and new ways of finding job vacancies involve direct telephone calls to employers, job agency offices, through personal contacts, scanning online job listings (which sometimes are vaguely described). Before the internet became widely used as a means for seeking jobs, candidates spent a lot of time using various methods to look for job openings.

Following are the traditional methods for seeking jobs:

- Job fairs.
- Putting up adverts in the mass media, for example, newspapers etc.
- Advertisements on the radio or tv.
- Through Management consultants.
- Placements through college or universities.
- Professional referrals.

These old methods of seeking jobs are very tedious, time consuming, lack quality and can sometimes even create confusion. In addition to that, the aspirants have to consider the cost and the amount of time to get to the information that they require. Also most of the companies or organisations are using the internet as a means to find their perfect candidate. But even in that approach there are some pitfalls.

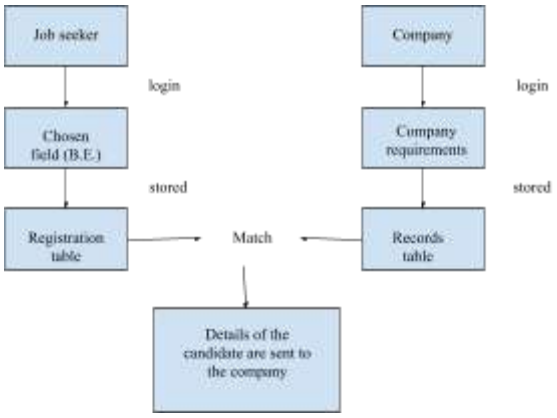
B. Importance of HireEasy: In this day and age, the internet is a goto place for any kind of search, especially job searches. Employers as well as candidates save time & money. But because the internet is open to each and everyone anyone can post a job advertisement, some of which are even a scam. Also the candidates can simply with the click of a button, apply to any job listing without giving a second thought, which in turn provides undesirable outcomes.

Using the benefits of machine learning and data from various candidates as well as placement records of various companies, the

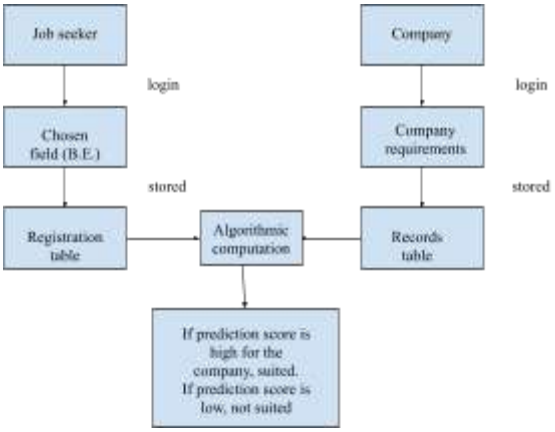
employers will be able to get their desired candidates within a short period of time and with reduced efforts. Also the candidate with better able to understand, with the help of his predicted result, which company is he/her eligible to apply.

[2] Several e-recruitment systems have been proposed with an objective to speed-up the recruitment process, leading to a better overall user experience. E-Gen system performs
analysis and categorization of unstructured job offers, i.e. in the form of unstructured text documents) as well as analysis and relevance ranking of candidates. CommOn framework [4] applies Semantic Web technologies in the field of Human Resources Management. In this framework, the candidate's personality traits, determined through an online questionnaire which is filled-in by the candidate, are considered for recruitment.

3. Existing system



4. Proposed system



5. Existing methodology and systems

The existing system for job recruitment includes traditional methods like employment agencies, advertising through newspapers, televisions and radios, college fairs etc. which are slow, stressful. With the advancement in technology, through algorithms it is possible to predict the suitability of a candidate for a job.

6. Proposed methodology

HireEasy is a php based web application that provides a predictability score by taking into account the data given by candidate at the time of registration (provided that it is correct) and the hiring record of companies in the past. Based on the score of predictability, the candidate will apply only to the companies where his predictability

score is higher. This will save the time of both the employer and the candidate.

7. Algorithm

HireEasy uses Support Vector Machine algorithm to generate the prediction score. [3] **Support-vector machines** (**SVMs**, also **support-vector networks**) are the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on the side of the gap on which they fall.

In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

When data are unlabelled, supervised learning is not possible, and an unsupervised learning approach is required, which attempts to find natural clustering of the data to groups, and then map new data to these formed groups.

The support-vector clustering algorithm, created by Hava Siegelmann and Vladimir Vapnik, applies the statistics of support vectors, developed in the support vector machines algorithm, to categorize unlabeled data, and is one of the most widely used clustering algorithms in industrial applications.supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting).

8. Future scope

The datasets of more companies can be used to provide better rate of prediction. Also the prediction result can be simplified by showing a graphical representation of the score.

9. Conclusion

The proposed system aims at providing a candidate a better chance of getting the job by using his predictability score and also to the employer in finding the right candidate which fulfils their requirements.

IoT based Smart Healthcare Monitoring System

Reeya Wakchaure, Snehal Rokade, Diksha Varma

and Prof. Suhas Lawand[3]

1Member, Pillai College of Engineering, New Panvel, Maharashtra – 410206,India

2Member, Pillai College of Engineering, New Panvel, Maharashtra – 410206,India

3Member, Pillai College of Engineering, New Panvel, Maharashtra – 410206,India

3Guide, Pillai College of Engineering, New Panvel, Maharashtra – 410206, India

Abstract: **The paper presents the design and implementation of an IoT based Health Monitoring System for emergency medical services using the android technology. The project can demonstrate collection and integration of IoT data which offers adaptable operation and cost sparing alternatives to both medicinal services experts and patients. Here, a patient can be examined utilizing a collection of lightweight sensor nodes for real time sensing and analysis of different fundamental parameters of patients. The proposed result of the project is to provide accurate and efficient clinical services to patients by associating and gathering information data through health status monitors which would comprise the patient's temperature, blood pressure, oxygen saturation and keeps the patient updated with his present status and full medical data. Therefore, patients will have high quality services on the grounds that the system supports medical staff by providing the current status of patients health by eliminating the manual data collection.**

Keywords : IoT, Alert system, Emergency app, Sensors, Temperature, Blood Pressure, Pulse Oximeter

I. Introduction

IoT could be an arrangement of reticulated computing gadgets, mechanical and computerized machines, objects, animals or the individuals who are provided with particular images and therefore the capacity to transfer knowledge over a system while not expecting human to human or human to pc interaction. Associate degree IoT framework comprises web empowered sensible gadgets that are implanted processors, sensors and communication hardware to accumulate, send and follow up on information they procure from their surroundings. A thing in the Internet of things can be an individual with a heart monitor implant, an animal with a biochip transponder, a car that has built-in sensors to alert the driver when tire pressure is low or some other characteristic or man-made item that can be assigned an Internet Protocol (IP) address and can transfer information over a system. In the care stream, sensible system technology brings about higher diagnostic tools to higher treatment and personal satisfaction for patients by simultaneously diminishing costs of public care systems. An indicator could be a gadget that recognizes and responds to some style of input

from the physical environmental factors. This particular input can be lightweight, heat, movement, moisture, pressure or anyone of a good range of various natural phenomena. The output is generally a symptom that is born-again to human legible show at the detector location or transmitted electronically over a system for reading or any procedure. Sensors are around for various types like LDR, Temperature indicator, inaudibility identifier, moisture sensor, vibration detector, gas detector and others. Emergency application is a mobile application that indicates doctors and members from the family with respect to the emergency state of the patient.

II. Literature Survey

This section reviews the existing recent literature work and provides insights in understanding the challenges and tries to find the gaps in existing approaches. Various computing techniques are applied in the Healthcare domain. The focus of the literature survey here is on the use of Internet of Things and Android systems in the healthcare domain. Health- care domain challenges are in improving research phases.

Internet of things (IoT) based smart health care system [1]

A BSN (Body Sensor Network) is a system intended to operate autonomously to associate various medical sensors and implants situated inside and/or potentially outside of the human body; which offers adaptable operation and cost sparing alternatives to both healthcare professionals and patients. This work illustrates the structure and implementation of a smart health monitoring framework. Here, a patient can be monitored using a collection of lightweight wearable sensor nodes for real time detecting and analysis of various indispensable parameters of patients. The devices seamlessly assemble and share the information with each other and furthermore store the information, making it possible to gather records and analyze data. The different sensors are placed at the respective locations on the human body and are connected to the Arduino board. The temperature sensor output from LM35 is converted to digital form with the assistance of ADC pins of the Arduino board. For the pulse rate sensor when the heart pumps blood through the blood vessels, the finger turns out to be marginally opaque and so less light arrives at the detector. With each heart beat the detector signal differs and this variation is converted into an electrical pulse. The pulse is also indicated by an LED which blinks on every heartbeat.

Body Temperature Measurement for Remote Health Monitoring System [2]

The goal of this project is to structure and develop a body temperature estimation gadget framework that can be monitored by the specialists in real time as well as history data via the web with an alarm/indication in occurrence of abnormalities. In the proposed health monitoring system, pulse rate and body temperature wireless remote sensors were developed, however this paper only focuses on body temperature wireless monitoring framework. The temperature sensors will send the sensed readings to a microcontroller using Xbee wireless correspondence. In order to transmit the continuous sensed information to the health monitoring database, wireless local area network (WLAN) has been utilized. An Arduino with Ethernet shield dependent on IEEE 802.11 standard has been used for this purpose. Test results from a group of volunteers shows the real-time temperature reading successfully monitored locally (at home) and remotely (at doctor's computer) and the readings are comparable to commercial thermometers.

IOT BASED PATIENT HEALTH MONITORING SYSTEM [3]

The fundamental thought of the designed framework is to persistently monitor the patients over the internet. The model comprises

LPC2148 Microcontroller,Temperature sensor(LM35), Heart Beat Sensor, Liquid Crystal Display(16x2), GPRS Modem, Piezo Electric Buzzer, Max232, Regulated Power Supply. In this system LPC2148 Microcontroller gathers the information from the sensors and sends the information through GPRS Protocol. The Protected information sent can be accessed anytime by the doctors by typing the corresponding unique IP address in any of the Internet Browsers at the end user device(ex: Laptop, Desktop, Tablet, Mobile phone). The Microcontroller is associated with GPRS Modem which provides information to doctor/caretaker when the heart rate goes irregular. During this time the buzzer turns on and alerts the caretaker. The user interface html webpage will automatically refresh for every 15 seconds subsequently patient health status is persistently sent to the specialist.

III. Proposed System

A. Proposed Architecture of Smart Healthcare Monitoring Systems

This can be contributed essentially to the improvement in the classification and recognition frameworks utilized in disease diagnosis which is able to provide information that guides medical specialists in early detection of lethal diseases and therefore, increment the survival rate of patients altogether. The application of android frameworks in the field of medical diagnosis is expanding gradually. The consequences of the study reinforce the idea of the application of IoT and android systems in early detection and outcome of diseases. Medicinal services and health of individuals is an important necessity of the human population. Heart Rate and body temperature are two significant vitals associated with the health of an individual. The ability to screen these two vital signs is key to guarantee legitimate healthcare is delivered early. In this paper, a system to monitor heart rate, body temperature and blood pressure of a user and alert the user when these values are anomalous is proposed. Patient Medical Emergency Alert System (PMEAS) essentially comprises two components, a wearable hardware unit and an android application. The wearable unit contains sensors to monitor the heart rate and body temperature of the user, which are displayed on an LCD screen.

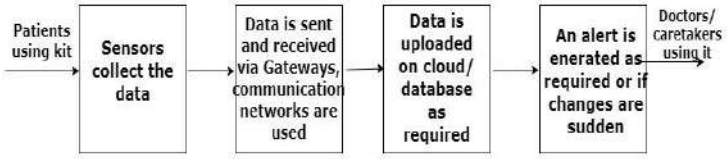


Fig. 1

B. Flowchart

The flow of the system is being specified as the detection is being performed by the sensors and the alert system eventually works as per the threshold value.

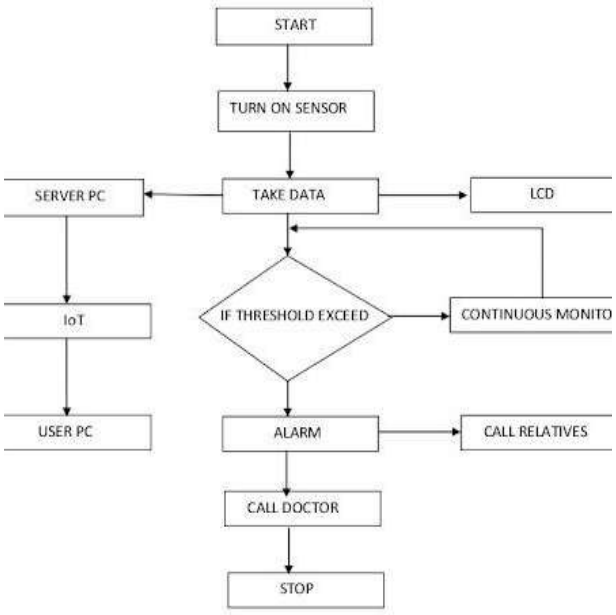


Fig.2

IV. Materials and Methods

A. Body Temperature

Body temperature (BT) is a consequence of the harmony between heat production and heat loss in the body, being its estimation crucial to avoid many elements defunctionalization because of high temperatures (e.g., proteins denature and lose function above certain temperatures)

Fig. 3

Lm35 is a temperature sensor that outputs an analog signal which is proportional to the instantaneous temperature. The output voltage can easily be interpreted to obtain a temperature reading in Celsius.

B. Blood Pressure Measurement

Blood pressure (BP) is considered as a crucial cardiopulmonary parameter, demonstrating the pressure exerted by blood against the arterial wall. BP provides roundabout information about the blood flow when the heart is contracting (systole) and relaxing (diastole) and can likewise demonstrate cellular oxygen conveyance. Ambulatory BP monitoring permits getting BP readings several times a day, which is ideal to monitoring high blood pressure (hypertension), probably one of the greatest threats to the global burden of illnesses, improving cardiovascular disease prediction. The gadget which quantifies this is called Sphygmomanometer. Here we use a wrist blood circulatory pressure monitor as shown in Fig 2.

Fig. 4

C. NodeMCU

The NodeMcu is an open-source firmware and development kit that helps you to prototype your IoT product with a couple Lua script lines.The Development Kit as appeared in Fig. 3 is depends on ESP8266, integates GPIO, PWM, IIC, 1- Wire and ADC across the board. ESP8266EX (potentially referred to as ESP8266) is a framework on-chip (SoC) which incorporates a 32-bit Tensilica microcontroller, standard advanced digital peripheral interfaces, antenna switches, RF balun, power amplifier, low noise receive amplifier, channels, filters and power management and execution modules into a small package.

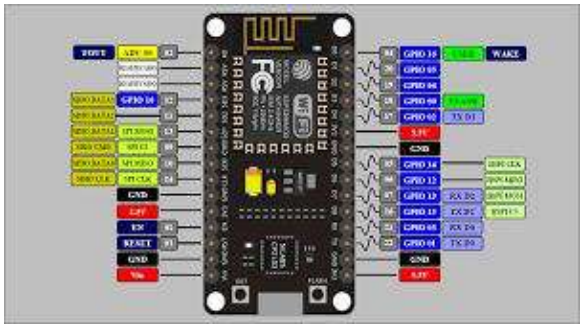


Fig. 5

D. Pulse Oximeter Sensor

Hardware Description Pulse oximetry is a basic technique to discover the proportion of haemoglobin. Oximeter quantifies the number of hearts beat per unit of time which is usually conveyed in bits per minute (Bpm). In the undertaking MCP6004 based pulse oximeter is designed and TCRT1000 reflective IR optical

sensor is utilized for photoplethysmography(PPG). UsingTCRT1000 improves the strategy since both emitter and detector are arranged side by side. This technique is used to evaluate heart rate since change in blood volume is synchronous to heart beat.

V. Implementation Algorithm of the proposed system

A. WiFi Connections

In this project, we are going to utilize ESP 12 as a WiFI module to create association with the internet. The ESP8266 WiFi Module is an independent SOC with integrated TCP/IP protocol convention stack that can provide any microcontroller access to your WiFi network. The ESP8266 is capable of either facilitating an application or offloading all Wi-Fi networking capacities from another application processor.

B. Working

Correspondence between ATMEGA 328 and Node MCU is done over the sequential serial port. ATMEGA328 sends information on the transmit pin and Node MCU receives it on the receive pin. The majority of the computation part is dealt by the ATMEGA 328 and then data is sent to the web server using the Node MCU. Node MCU is Wifi empowered which can communicate with the Wi-Fi Source for the connection and then provides the information to the web servers via the Get requests. Also, for the correspondence between them both should have the same Baud rates.

The values accepted by the server are then stored in the database. The values in the database are analysed thoroughly and are provided with ideal thresholds. These thresholds provide a basis to determine anomalies in readings. If the values of a particular health parameter exceed the threshold, which means a potential threat, an alert SMS is sent to the individuals on the priority list of the patient's contacts including the assigned doctor through the SMS Gateway. All the sensors are connected to the printed circuit board. The values stored in the database are displayed in the emergency app continuously between time intervals. The picture depicting the constant readings displayed in the app are shown in Fig. 7

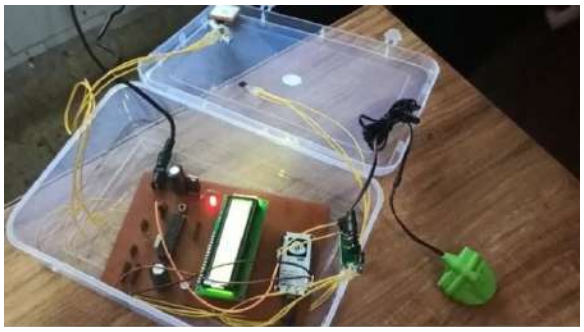
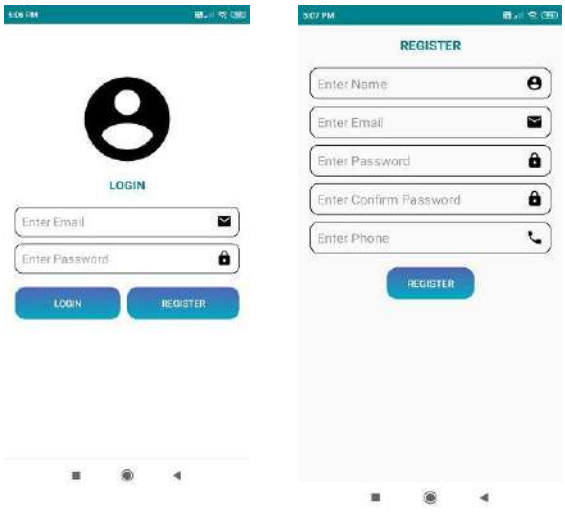


Fig. 6

C. Alert system.

The wireless monitoring of patients has an incredible effect in the field of medicine. With the assistance of micro sensors which are integrated into wireless communication networks, the physiological parameters of the patient can be remotely accumulated and observed utilizing traditional medical instruments can be avoided. In this project the monitoring of the patient is undertaken by the doctor consistently without actually visiting the patient. Here, we make use of various IoT sensors that detect like pulse rate, temperature, blood pressure of the patient. These detected signals are transmitted to update the data constantly. The readings are sent to the individual on the priority list wirelessly. Consequently, the specialist can monitor the patient from anywhere. This framework has the capability of providing real time monitoring and also improvements of SMS. It sends the data to the control unit for additional processing and estimations are displayed on the application. It then proceeds to alert by sending SMS, SMS will be sent to the cell phone of the individuals in the priority list, if and only if the threshold value is maximally surpassed.

D. Emergency App



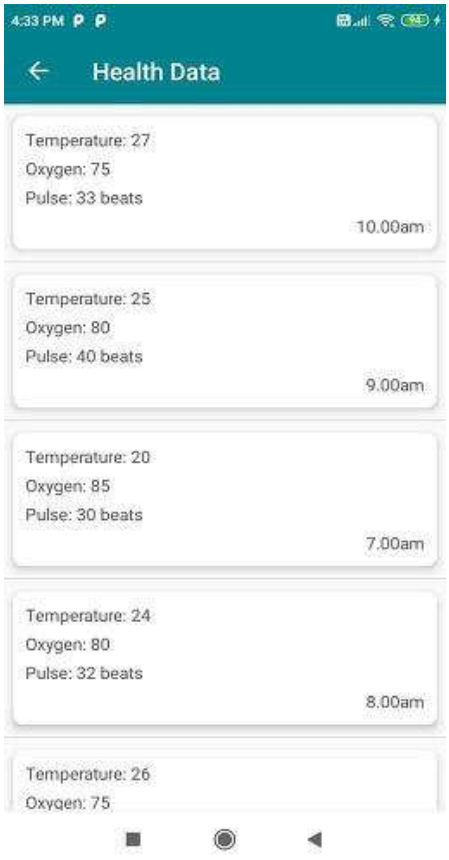


Fig. 7

VI. Conclusion and Future Scope

The objective of this project has been successfully achieved. Body temperature, oxygen saturation and blood pressure measurements for remote health monitoring have been designed and developed. The system provides reliable measurements and is very user friendly. The system successfully alerts the guardians in case of abnormal readings.

The device and the system can be improved in terms of sizing and integration between more measurement devices with the existing set-up, for example electrocardiography (ECG).

References

[1] Vikas Vippalapalli and Snigdha Ananthula "Internet of things (IoT) based smart health care system" International conference on Signal Processing, Communication, Power and Embedded System (SCOPES)-2016

[2] Hasmah Mansor , Muhammad Helmy Abdul Shukor, "Body Temperature Measurement for Remote Health Monitoring System" Proc. of the IEEE International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA) 26-27 November 2013, Kuala Lumpur, Malaysia

[3] Prof. M.P. Sardey,Desale Pratik B. "IOT BASED PATIENT HEALTH MONITORING SYSTEM" International Journal of Advance Engineering and Research Development Volume 4, Issue 6, June -2017

[4] Sarfraz Fayaz Khan Dept. ofMIS, "Health Care Monitoring System in Internet of Things (loT) by Using RFID "2017 the 6th International Conference on Industrial Technology and

Management

[5] K. Natarajan, B. Prasath, P. Kokila "Smart Health Care System Using Internet of Things" Journal of Network Communications and Emerging Technologies (JNCET) www.jncet.org Volume 6, Issue 3, March (2016) ISSN: 2395-5317 ©EverScience Publications 37

[6] Raghavendra K K, Sharanya P S, Shaila Patil "An IoT Based Smart Healthcare System Using Raspberry Pi" International Journal of Research and Scientific Innovation (IJRSI) | Volume V, Issue VI, June 2018 | ISSN 2321–2705

[7] Sneha N. Malokar, Samadhan D. Mali, "An IOT Based Health Care Monitoring System-A Review" International Journal of Innovative Research in Computer and Communication Engineering

[8] Melisa Pereira, Nagapriya Kamath K "A Novel IoT Based Health Monitoring SystemUsing LPC2129" 2017 2nd IEEE International Conference On Recent Trends in Electronics Information & Communication Technology (RTEICT), May 19-20, 2017

[9] "An IoT-Based E-Health Monitoring System Using ECG Signal" Maryem Neyja, Shahid Mumtaz, Kazi Mohammed Saidul Huq, Sherif Adeshina Busari, 978-1-5090-5019-2/17/$31.00 ©2017 IEEE

E-learning Package for Grape & Disease Analysis

Akash Pimpalkar[1], Pradumna Patki[2], Sonali Patil[3], Dr.Sushopti Gawade[4]

[1][2][3]B.E students, Department of Information technology, Pillai College of Engineering, New Panvel, Maharashtra – 410206

[4]Professor,Department of Computer Engineering, Pillai College of Engineering, New Panvel, Maharashtra – 410206

Abstract— Classification of grape leaf disease is the main purpose to prevent the losses and quality of the agricultural product. In India, grapefruit crops are extensively grown. So disease detection and classification of the grape leaf is very crucial for sustainable agriculture. It's not possible for farmers to continuously observe grape disease manually. It requires excessive processing time, a large amount of work, and some expertise in the grape leaf diseases. To detect and classify the grape disease we need a fast automatic process so we use CNN technique. It involves the following steps in that, image acquisition, image pre-processing, features extraction, and neural network-based classification. The developed algorithm's efficiency can successfully detect and classify the examined disease with an accuracy of 91%. This paper is proposed to benefit in the detection and classification of grape leaf disease using CNN(Convolutional neural network).

Keywords: Image processing, detection, classification, neural network, CNN.

1. INTRODUCTION

India is an agricultural country.70% of the population depends on agriculture. A farmer has a wide range of diversity to select proper fruit and vegetable crops. Plant disease is gaining importance as it can cause a vital reduction in both qualities & gaining quality of the agricultural product. So, research on the detection of plant disease is gaining attention nowadays, which may prove useful in monitoring large fields and thus automatically detection symptoms as they appear on the plant. Grapes (Vitis vinifera)are a major fruit crop in India. Grapes are popularly consumed as fresh fruit in India. It is also used for producing raisins, wine, juice, juice concentrate, squash, beverages, jams, and marmalades.

Grapefruit enjoys a pre-eminent status among all cash crops in a country and is a principal raw material for the flourishing wine industry. It also provides livelihood to about 65 million people and is an essential agricultural commodity providing remunerative income to millions of farmers in developed as well as in the developing country. About 60% of grapes cultivated in India are under rainfed condition. Water stressed seed or plant will cause reduced growth leading to low yield as well as exposure to disease Due to disease on the plant there is a loss of 10-30 % of the crop. Farmers do the naked eye observation and judge the diseases by their experience. But this is not an accurate and precise way. Sometimes farmers want to call the experts for identifying the diseases but this also a time-consuming way. Most of the disease on the plant is on their leaves and on the stem of the plant. The diseases are categorized into viral, bacterial, fungal,
diseases due to insects, rust, nematodes, etc. on the plant. Early detection of diseases is a major challenge in horticulture/agriculture science.
Computer vision systems will help to tackle the problem. Computer vision systems developed for agricultural applications, specifically detection of weeds, sorting of fruits in fruit processing, classification of grains,

identification of food products in food processing, medicinal plant identification, etc. In all these techniques, digital images are taken in a given domain using digital cameras and image processing techniques are implemented on these images to extract useful features that are necessary for additional analysis.

2. LITERATURE SURVEY

Agricultural plant leaf disease detection using image processing [1]. This paper describes that there are mainly four steps in developed processing scheme, out of which, first one is, for the input RGB image, a color transformation structure is formed because this RGB is used for color formation and modified or converted image of RGB, that is, HSI is used for the color descriptor. In the second step, by using a threshold value, green pixels are masked and removed. In third, by using the pre-computed threshold level, removing green pixels and masking is done for the segments that are obtained first in this step, while the image is segmented and in the last or fourth main step, the segmentation is done.

Detection of unhealthy regions of plant leaves and classification of plant leaf diseases using texture features[2]. This paper demonstrates the disease identification process include some steps out of which four main steps are as follows: first, for the input RGB image, a color transformation structure is taken and then using a particular threshold value, the green pixels are masked and removed, which is further followed by segmentation process, and for getting useful segments the texture statistics are computed. At last, the classifier is used for the features that are extracted to classify the disease. The robustness of the suggested algorithm is proved by using experimental outcomes of about 500 plant leaves in a database.

Image processing techniques for the detection of leaf disease [3]. The state of the art review of different methods for leaf disease detection using image processing techniques is presented in paper[3]. The present methods studies are for boosting throughput and reduction subjectiveness which happens due to naked eye observation through which identification and detection of plant diseases are done.

Advances in image processing for the detection of plant diseases[4]. According to [4] histogram matching is used to identify plant disease. In plants, the disease appears on leaves therefore the histogram matching is done on the basis of edge detection technique and color feature. Layers separation technique is used for the training process which encompasses the training of these samples which separate the layers of RGB image into red, green, and blue layers and edge detection technique that identifies edges of the layered images. Spatial Gray Level Dependence Matrices are applied for improving the color co-occurrence texture analysis method.

A survey of plant leaf disease detection techniques[5]. The author of the paper explains the different symptoms for different types of plant diseases along with the technology that is being developed to make this process easy. This paper also highlights the different types of techniques that are currently available and are used for the purpose of disease detection as early as possible.

Usability improvement with crop disease management as a service [6]. This paper gives the reader information about how the digital space in the technology can be used to detect and manage different plant diseases. The major focus of this paper is to improve usability services in the agriculture sector by providing better tools.

Detection of plant leaf diseases using image segmentation and soft computing techniques[7]. This paper explains an algorithm for image segmentation technique which is developed solely for the purpose of automatic plant disease detection. The paper also covers surveys on different disease classification techniques that can be used for plant leaf disease detection.

Plant disease recognition based on image processing techniques[8]. The author of this paper explains the multiple linear regression techniques in depth. This paper also highlights how multiple linear regression can be used to make the system more accurate, efficient and intelligent

3. PROPOSED SYSTEM ARCHITECTURE

The main objective behind creating such an intelligent system is to avoid heavy agricultural losses that are faced by the farmers every year. The system architecture is designed in such a way that efficiency is maximum. The system basically uses the Convolution neural network algorithm to compare the image uploaded with the image present in the database. The comparison is done on the basis of the features extracted using different image processing techniques. The steps for the above-mentioned process are given below along with their specific explanation and how this helps to solve the aforementioned problem.

After study several literature reviews, there is a need to develop a real-time system that will efficiently use to detect diseases on the grape plant. The task of plant disease classification and classification is of greater importance in the field of agriculture. Therefore, developing automated techniques for plant disease classification has got much interest in the field of research nowadays. To diagnose the disease, an image processing system has been developed to automate the identification and classification of various diseases.

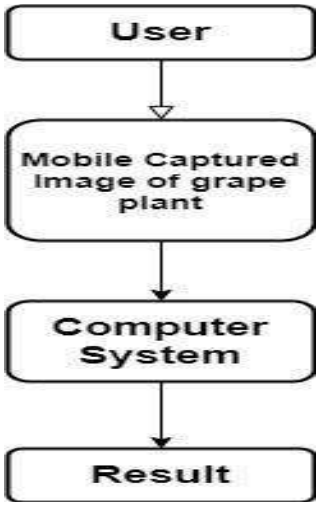


Fig 3.1: Overall structure of the system

We introduce an image-processing-based solution for automatic leaf disease detection and classification. We test our solution on three diseases which affect the plants; they are Black rot, Black Measles, and Leaf blight(Isariopsis_Leaf_Spot). Firstly, the digital images are taken from the environment using a digital camera. Then image-processing techniques are used to the acquired images to extract useful features that are required for further analysis. After that, several analytical distinguishing techniques are used to classify the images according to the specific problem at hand fig.2 depicts the basic procedure of the proposed vision-based detection algorithm in this research.

3.1 Description of the overall system

1) **User**: Users who want to know the kind of disease on the grape leaf will capture the image and sent it to the computer
system.

2) **Computer System**: The actual trained model which classifies the disease is stored in the PC. The disease affected leaf to be tested using the CNN model in the PC which gives the more accurate results and displays the result.

3.2 System Algorithm

Basic steps for describing the proposed system are as follows:

1. Acquire an image from a farmer.
2. Give image input to the system.
3. Image Pre-processing
4. Segmentation

5. Extract the features.
6. Apply classifier.
7. Get result.

3.3 Software architecture

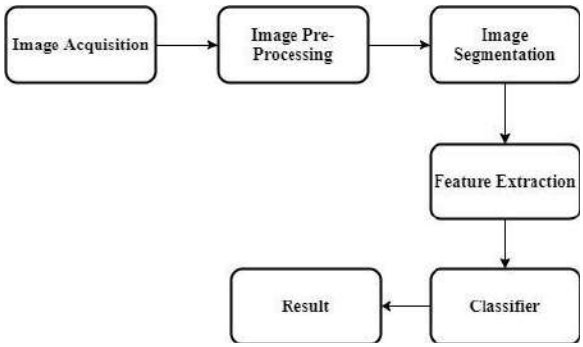


Fig 3.2: Steps for leaf disease detection

1) Image acquisition:

Firstly, the images of several leaves taken using a digital camera with the required resolution for better quality. After that sample images are obtained or collected from the farm of grape using different mobile cameras with different resolutions that are used to train the system. Collected images cover the healthy leaf as well as affected leaf by different diseases like black rot, Black Measles, and Leaf blight(Isariopsis_Leaf_Spot), etc.

2) Image pre-processing:

In the second step, this image is pre-processed to enhance the image Pre-processing covers color conversion, histogram, and histogram equalization. Color conversion and histogram equalization are used to improve the quality and clarity of images. The histogram equalization enhances the contrast of images by changing the intensity values. Colour conversion of RGB to Gray image is done using the following equation:

f (X) = r*0. 2989+ g*0. 5870+b*0.114

3)Segmentation:

It means the representation of the image in a more meaningful and easy to analyze. In segmentation, the digital image is partitioned into multiple segments can be defined as super-pixels.

4)Feature extraction:

Extracting the important data from the input image is the process of feature extraction. Also converting the input data into the set of features is called feature extraction. There are several types of features of leaf images such as color, texture, shape, and edges, etc. so in this proposed system color and texture features are selected to get good results and accuracy.

5) Classification:

The classification technique is used for both the training and testing process. This is the last and final step of the system. The features obtained from training leaves are matched with features extracted from testing leaves. Then the images are classified or identified based on the matched features. So the CNN(Convolutional neural network) model is used for the classification of leaf disease. This excludes the need for manual feature extraction. The features are not trained! They're learned while the network trains on a set of images. This delivers deep learning models extremely accurate for computer vision tasks. CNNs learn feature detection through tens or hundreds of hidden layers. Each layer enhances the complexity of the learned features.

4. RESULT

The goal of the proposed approach was to analyze and predict the disease of the grape leaf.

Fig 4.1: Web GUI

Fig 4.2: Web GUI after giving input(Black rot)

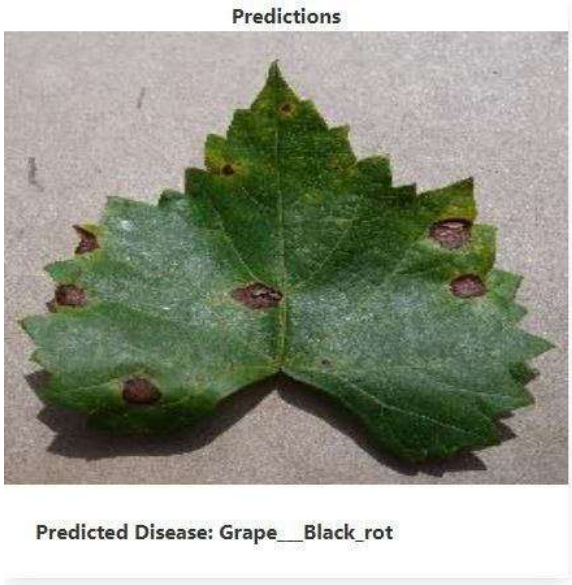


Fig 4.3: After prediction

Fig 4.4: Predictions with remedies

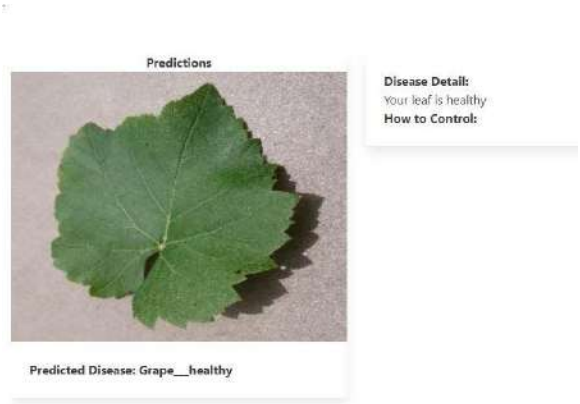


Fig 4.5: prediction(Healthy Leaf)

5. CONCLUSION

We have developed leaf disease detection and analysis system with the help of CNN model which is capable of detecting disease on leaves. A set of features was chosen to be extracted using the feature extraction phase, and those features were collected in the feature database, which is designed for this purpose. The captured leaf image parameters were correlated with the parameters of healthy leaf and disease was identified. The classification rate is above 93.32% In the experiments, the purposes for misclassification of the plant disease are concluded as follows: the indications of the texture of diseased plant leaves vary at the beginning. To improve the plant disease classification rate at various stages, we need to grow the training samples and extract the effective features from leaf texture.

ACKNOWLEDGMENT

It is our privilege to express our sincerest regards to our supervisor Dr. Sushopti Gawade 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 to our Head of the Department Dr. Satishkumar Varma and our Principal Dr. Sandeep M. Joshi for encouraging and allowing us to present this work.

6. REFERENCES

[1]Sujatha R* , Y Sravan Kumar and Garine Uma Akhil,"Leaf disease detection using image processing ", Journal of Chemical and Pharmaceutical Sciences,January - March 2017.

[2] Vijai Singh, A.K. Misra, "Detection of plant leaf diseases using image segmentation and soft computing techniques", Information Processing In Agriculture 4 (2017) 41–49

[3]Naikwadi Smita, Amoda Niket. Advances in image processing for detection of plant diseases. Int J Appl Innov Eng Manage 2013;2(11).

[4]Rathod Arti N, Tanawal Bhavesh, Shah Vatsal. "Image processing techniques for detection of leaf disease". Int J Adv Res Comput Sci Softw Eng 2013;3(11).

[5]Komal Raikar, Sushopti Gawade, Varsha Turkar. "Usability improvement with crop disease management as a service". 2017 International Conference on Recent Innovations in Signal processing and Embedded Systems (RISE).

[6]Nivedita.R.Kakade , Dnyaneswar. D.Ahire. "Real time grape leaf disease detection", IJARIIE-ISSN(O)-2395-4396 (Vol-1 Issue-4 2015).

Sentence Similarity System

Shubhangi Chavan
Department of Computer Engineering,
Pillai College of Engineering,
New Panvel, India
srathod@mes.ac.in

Darshan Kadam
Department of Information Technology,
Pillai College of Engineering,
New Panvel, India
darshankadam8@gmail.com

Nayan Joshi
Department of Information Technology,
Pillai College of Engineering,
New Panvel, India
nayan.joshi98@gmail.com

Sahil Kadu
Department of Information Technology,
Pillai College of Engineering,
New Panvel, India
sahilkadu12@gmail.com

***Abstract*— The measure of how similar the given sentences are can be called as Sentence similarity, which plays an important role in text-related research and application in area such as text-mining. In this system we Pre-process the given sentences in a bag of words using Tokenization, Stemming and other Natural language techniques. Then we apply syntax similarity techniques and semantics similarity techniques. The syntax similarity technique finds the grammatical syntax similarity between sentences. The Semantic similarity technique finds the Semantic similarity between words, it creates a relationship between words and sentences through the meanings of the words. The technique used to calculate Syntactic similarity are Cosine similarity, Word order similarity and Jaccard similarity. In Cosine similarity we find the cosine similarities between sentences, in Word order similarity we find the Intersection word set of them which contains common words between the sentence. The sentence similarity is used in Plagiarism detection system, Question-Answering system, etc.**

Keywords*— Semantic similarity, syntactic similarity, WordNet, Hindi similarity, Text similarity.

I. Introduction

Natural Language Processing, usually shortened as NLP, is a branch of artificial intelligence that deals with the interaction between computers and humans using the natural language. The ultimate objective of NLP is to read, decipher, understand, and make sense of the human languages in a manner that is valuable. Most NLP techniques rely on machine learning to derive meaning from human languages. NLP is commonly used in many applications such as.

- Language translation applications such as Google Translate
- Personal assistant applications such as Google Assistant, Siri, Cortana, and Alexa.
- And many more...

NLP entails applying algorithms to identify and extract the natural language rules such that the unstructured language data is converted into a form that computers can understand. When the text has been provided, the computer will utilize algorithms to extract meaning associated with every sentence and collect the essential data from them. Sometimes, the computer may fail to understand the meaning of a sentence well, leading to obscure results. In current times the Sentence Similarity measures are used more and needed in the Text-based Research and other areas. Some similarity measure calculates the similarity between 2 sentences, thoroughly using Word-net semantic dictionary. The Sentence Similarity is the one of the core functions in NLP tasks such as Paraphrase detection, etc. Given the two sentences , the task of calculating the similarity is defined how similar is the meaning of two sentences. The higher the similarity, the more similar the meaning of two sentences are.

II. Proposed Model

A. Overview

This section presents an Overview and Description of techniques used for the system. The Semantic similarity is calculated using Word-net as the corpus.

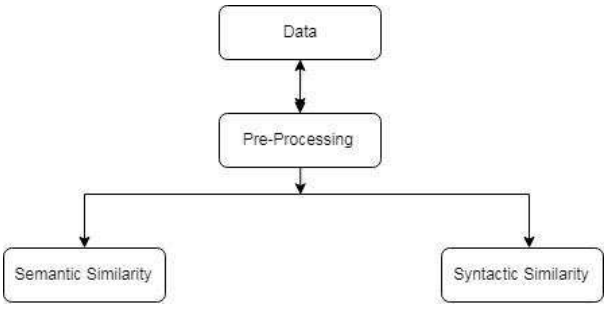


Figure 1 Overview of System

This system checks the similarity between sentences, giving a similarity output which presents how much percent of the sentences are similar. This System can be used to detect plagiarized documents.

Processing Steps :

1. **Tokenization:** A given Sentence is divided into array/list of separate words called Tokens.
2. **Lemmatization:** It brings the word to its base form using the proper vocabulary.

3. **Stop-Word Removal:** The removal of the Stopwords (such as “is”,”the”.etc) is called Stop Word Removal.
4. **Semantic Similarity:** Similarity returns a score denoting how similar two word or sentence senses are, based on some measure that connects the senses in is-a taxonomy.
5. **Syntactic Similarity:** Syntax similarity is a measure of the degree to which the word sets of two given sentences are similar on the basis of their syntax.

B. Existing System

The presented System Architecture is using the syntactic approach and semantic with only one syntactic technique:

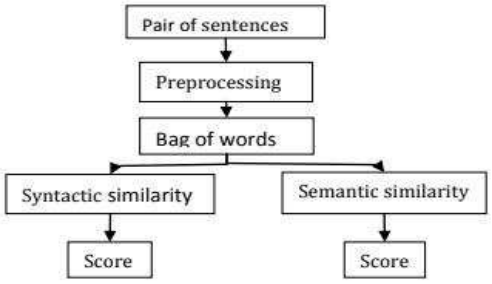


Figure 2 Existing System[1]

This approach uses only both Syntactic and Semantic way of the similarity measure. In the Semantic approach the techniques used :

1. Wu-Palmer similarity
2. Feature based measure
3. Short Path distance

The Syntactic approach only uses a single technique for measurement between text :

1. Cosine Similarity

The system checks the only uses one similarity technique for syntactic level of similarity calculation. This makes the system not so accurate on the level of Syntactic similarity.[1]

III. Proposed System

As discussed above, architecture does not have the more accurate syntactic score and all the scores aren’t calculated in a single score for the system.

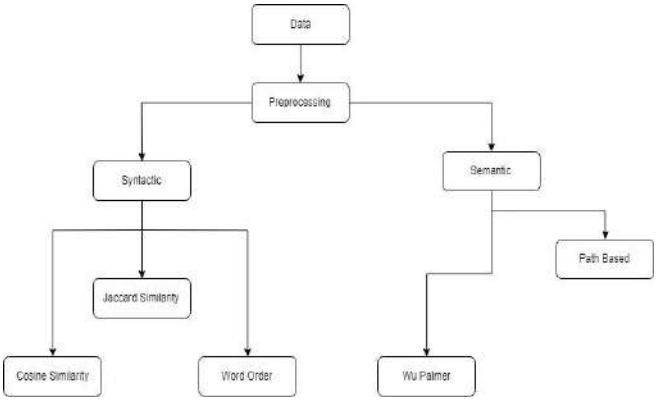


Figure 3 Proposed System

In the proposed system, the data is first passed through the Pre-Processing stage where the data get processed using the NLP (Natural Language Processing) techniques. Then the processed data is created in BOW (Bag of words). This processed data is used in different Syntactic and Semantic techniques.

The techniques gives a score as per the algorithm presented. This score is calculated as per a semantic level and syntactic level as a single score for each level. This score is then presented to the user.

A. Hardware and Software Specifications

The experiment setup is carried out on a computer system which has the different hardware and software specifications as given in Table 1 and Table 2 respectively.

TABLE I. Hardware details

Processor	1.6Ghz Intel/Amd
Graphics	512Mb
Ram	4 GB

TABLE II. Software details

Programming Language	Python 3.6, Html, Css
IDE	IDLE/Pycharm
Operating system	Windows 7 and up.
Database	Mysql
Browser	Chrome

Project Inputs and Outputs

B. Input Details

The project takes two sentences as input. The input is in Hindi language, this input is processed through in back-end for display of results.

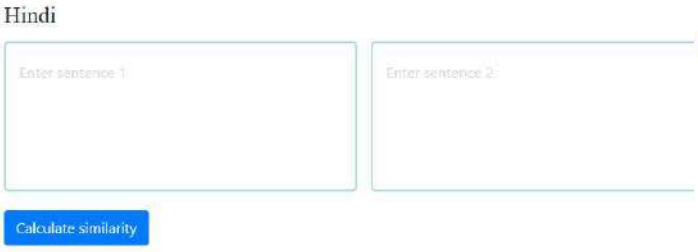


Figure 4 Input for the system

C. Data Processing and Evaluation

The data is given from the is Input is Pre-processed using the Pre-processing techniques. The data is translated from the multiple language to English as the corpus for other languages is not much developed.

In Pre-Processing we use tokenization, Stemming, and Stop Word removal. The Pre-Processed data is used for calculation of sentence similarity.

In Sentence Similarity, we apply Syntactic and Semantics Similarity. The Syntactic Similarity approach uses the following algorithms:

1. Cosine Similarity
2. Jaccard Similarity
3. Word-Order Similarity

1. Cosine Similarity

We convert the preprocessed data into vector array then find the cosine angle between the vector array of he both sentences.[1]

2. Jaccard Similarity

In Jaccard Similarity, we find the union of data between the 2 sentences and average it out with the total words in both sentences.[4]

3. WordOrder Similarity

We take bigrams of the sentences and hash them then find the similarity of the hash between the both sentences.[8]

The Semantic approach uses the following algorithm:

1. Wu- Palmer
2. Path Based

1. Wu - Palmer

In this we use wordnet corpus to get synsets of the provided words and then use wu palmer to find the closest matching synset. Then create array of there synset for both sentences then take their dot product for determining the similarity.[1]

2. Path Based

In this we use wordnet corpus to get synsets of the provided words and then use path based nltk algorithm to find the closest matching synset. Then create array of there synset for both sentences then take their dot product for determining the similarity.

```
def tokenize(self):
    '''done'''
    if not self.sentences:
        self.generate_sentences()

    sentences_list=self.sentences
    tokens=[]
    for each in sentences_list:
        word_list=each.split(' ')
        self.tokens=self.tokens+word_list
    self.hyphenated_tokens()

# def print_tokens(self,print_list=None):
#         '''done'''
#         if print_list is None:
#             for i in self.tokens:
#                 print (i.encode('utf-8'))
#         else:
#             for i in print_list:
#                 print (i.encode('utf-8'))


def generate_stem_words(self,word):
    suffixes = {
1: [u"ो",u"े",u"ू",u"ु",u"ी",u"ि",u"ा"],
2: [u"कर",u"ाओ",u"िए",u"ाई",u"ाए",u"ने",u"नी
3: [u"ाकर",u"ाइए",u"ाई",u"ाया",u"ेंगी",u"ेंगा",
4: [u"ाएगी",u"ाएगा",u"ाओगी",u"ाओगे",u"एंगी",u"
5: [u"ाएंगी",u"ाएंगे",u"ाऊंगी",u"ाऊंगा",u"ाइयाँ",u"
}
    for L in 5, 4, 3, 2, 1:
        if len(word) > L + 1:
            for suf in suffixes[L]:
                if word.endswith(suf):
                    return word[:-L]
    return word
```

Figure 5 Pre-processing of Data

```
def cosine(text1,text2):
    WORD = re.compile(r'\w+')

    def get_cosine(vec1, vec2):
        intersection = set(vec1.keys()) & set(vec2.keys())
        numerator = sum([vec1[x] * vec2[x] for x in intersection])
        sum1 = sum([vec1[x] ** 2 for x in vec1.keys()])
        sum2 = sum([vec2[x] ** 2 for x in vec2.keys()])
        denominator = math.sqrt(sum1) * math.sqrt(sum2)
        if not denominator:
            return 0.0
        else:
            return float(numerator) / denominator

    def text_to_vector(text):
        words = WORD.findall(text)
        return Counter(words)


    vector1 = text_to_vector(text1)
    vector2 = text_to_vector(text2)
    cosine = get_cosine(vector1, vector2)
    print('Cosine:', cosine)

    return cosine
```

```
def wordorder(s1,s2,n = 2):
    tokens = [token for token in s1.split(" ") if token != ""]
    output1 = list(ngrams(tokens, n))
    tokens = [token for token in s2.split(" ") if token != ""]
    output2 = list(ngrams(tokens, n))

    count = 0
    hA = []
    hB = []
    for i in output1:
        hA.append(hash(i))

    for i in output2:
        hB.append(hash(i))

    for i in range(len(hA)):
        for j in range(len(hB)):
            if(hA[i]==hB[j]):
                count = count + 1
    sim1 = count/len(hA)
    sim2 = count/len(hB)
    return ((sim1+sim2)/2)
```

```
def jaccard(x, y):
    x = set(x)
    y = set(y)

    xy = x.union(y)
    yx = y.intersection(x)

    final = len(yx)/len(xy)
    print('FINAL', final)
    return final
```

Figure 6. Syntactic Similarity

```
def pb(s1, s2):
    tokenized = sent_tokenize(s1)
    for i in tokenized:
        wordsList = nltk.word_tokenize(i)
        wordsList = [w for w in wordsList if not w in stop_words]
    tokenized = sent_tokenize(s2)
    for i in tokenized:
        wordsList2 = nltk.word_tokenize(i)
        wordsList2 = [w for w in wordsList2 if not w in stop_words]
    s1 = set(wordsList)
    s2 = set(wordsList2)
    S = s1.union(s2)
    slen = len(S)
    v1 = []
    v2 = []
    for i in S :
        if(i in s1):
            v1.append(1)
        else:
            x = cmp(i,s1)
            if(x>0.25):
                v1.append(x)
            else:
                v1.append(0)
    for i in S :
        if(i in s2):
            v2.append(1)
        else:
            x = cmp(i,s2)
            if(x>0.25):
                v2.append(x)
            else:
                v2.append(0)
    def dot(K, L):
        return sum(i[0] * i[1] for i in zip(K, L))
    p = dot(v1,v2)
    return(p/slen)
```

```
def wp(s1, s2):
    tokenized = sent_tokenize(s1)
    for i in tokenized:
        wordsList = nltk.word_tokenize(i)
        wordsList = [w for w in wordsList if not w in stop_words]
    tokenized = sent_tokenize(s2)
    for i in tokenized:
        wordsList2 = nltk.word_tokenize(i)
        wordsList2 = [w for w in wordsList2 if not w in stop_words]
    s1 = set(wordsList)
    s2 = set(wordsList2)
    S = s1.union(s2)
    slen = len(S)
    v1 = []
    v2 = []
    for i in S :
        if(i in s1):
            v1.append(1)
        else:
            x = fwp(i,s1)
            if(x>0.25):
                v1.append(x)
            else:
                v1.append(0)
    for i in S :
        if(i in s2):
            v2.append(1)
        else:
            x = fwp(i,s2)
            if(x>0.25):
                v2.append(x)
            else:
                v2.append(0)
    def dot(K, L):
```

Figure 7 Semantic Similarity

D. Output Details

The output is displayed on a Web Application using a Bar Chart. This chart displays the similarity of the given data with "0" being the lowest and "1" being the highest similarity of the sentences provided as input.

Figure 8 Input

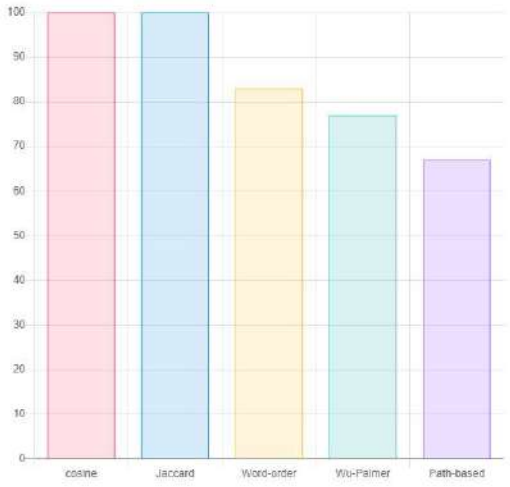


Figure 9 Output

IV. Conclusion

In this paper, the study of different Natural Language Processing techniques is presented. The different Preprocessing techniques such as Stop Word Removal, Tokenization, Stemming are explained. Different semantic measures such as Cosine similarity, Word Order, Jaccard Similarity are explained. The different syntactic approaches like Wu Palmer and Path Based Similarity are also described.

The comparative study of various techniques mentioned above is presented in this report. The performance measures the accuracy of the similarity of the given different data. The applications of this domain is identified and presented. It overcomes the limitations of previous system on the syntactic approach, which can give us more accurate results.

V. Future Scope

The accuracy of the whole system can increased by applying a Machine Learning model, which can be trained with large dataset consisting of various sentences. The model can be applied for each algorithm in the system to get more accurate results.

The semantic similarity score between sentences can be further improved by adding more semantic similarity algorithms in the system.

References

[1] Pantulkar Sravanthi, Dr. B. Srinivasu, “Semantic Similarity Between Sentences”. International Research Journal of Engineering and Technology (IRJET), 2017.

[2] Atoum, Issa & Otoom, Ahmed. (2016). Efficient Hybrid Semantic Text Similarity using Wordnet and a Corpus. International Journal of Advanced Computer Science and Applications. 7. 10.14569/IJACSA.2016.070917.

[3] H. Ruan, Y. Li, Q. Wang and Y. Liu, "A Research on Sentence Similarity for Question Answering System Based on Multi-feature Fusion," *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, Omaha, NE, 2016, pp. 507-510.

[4] Sarkar, Sandip & Saha, Saurav & Bentham, Jereemi & Pakray, Dr. Partha & Gelbukh, Alexander. (2016). NLP-NITMZ@DPIL-FIRE2016: Language Independent Paraphrases Detection.

[5] Sneha B., Mohit D., Zorawar Singh V. (2016) Comparison of Different Similarity Functions on Hindi QA System. In: Satapathy S., Joshi A., Modi N., Pathak N. (eds) Proceedings of International Conference on ICT for Sustainable Development. Advances in Intelligent Systems and Computing, vol 408. Springer, Singapore

[6] Sultan, M.A., Bethard, S. and Sumner, T., 2015. DLS $@ $ CU: Sentence Similarity from Word Alignment and Semantic Vector Composition. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015) (pp. 148-153).

[7] Hatzlvassiloglou, V., Klavans, J.L. and Eskin, E., 1999. Detecting text similarity over short passages: Exploring linguistic feature mbinations via machine learning. In 1999 Joint SIGDAT conference on empirical methods in natural language processing and very large corpora.

[8] Elkhidir, M., Ibrahim, M.M., Khalid, T.A., Ibrahim, S. and Awadalla, M., 2015, September. Plagiarism detection using free-text fingerprint analysis. In 2015 World Symposium on Computer Networks and Information Security (WSCNIS) (pp. 1-4). IEEE.

[9] Jingling, Z., Huiyun, Z. and Baojiang, C., 2014, November. Sentence similarity based on semantic vector model. In 2014 Ninth International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (pp. 499-503). IEEE.

[10] Gupta, D., Vani, K. and Singh, C.K., 2014, September. Using Natural Language Processing techniques and fuzzy-semantic similarity for automatic external plagiarism detection. In 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 2694-2699). IEEE.

[11] Abdi, A., Idris, N., Alguliyev, R.M. and Aliguliyev, R.M., 2015. PDLK: Plagiarism detection using linguistic knowledge. Expert Systems with Applications, 42(22), pp.8936-8946.

[12] Joshi, N., Kadam, D., Kadu, S. and Chavan, S., Survey on Sentence Similarity System.

Named Entity Recognition Using Syntactic Parsing For Hindi Language

Prem Thamarakshan [#1], Raj Paliwal [#2], Amit Shukla[#3], Shubhangi Chavan[#4]

[#1,2,3] Student , [#4] Professor
[#1,2,3,4] Department of Information Technology,
University of Mumbai,
Pillai College of Engineering, New Panvel, Maharashtra, India

---

Abstract - *NLP is a branch of artificial intelligence that deals with analyzing, understanding and generating the languages that humans use naturally so as to interpret with computers using natural human languages instead of computer languages. NER is the task of identifying named entities in a given text and distinguishing them based on their entity type. This paper discusses various techniques and models that have been discovered and are used for this process. It also provides analysis on how effective are these techniques and models in the process of Named Entity Recognition (NER). It also proposes a designed system which when implemented provides good accuracy.*

Key Words: **Hidden Markov Model, Rule based Approach and List Look Up Approach, Joint Parsing, NER identification, POS tagging.**

I. INTRODUCTION

Natural language processing is the proficiency of a computer system or program to understand and interpret human language. It is a component of computer science, linguistics and artificial intelligence. The development of NLP applications is challenging because computers traditionally require humans to tell them in terms of programming language that is precise, unambiguous and highly structured, or through a limited number of clearly enunciated voice commands. Human spoken language, however, is not always scrupulous -- it often possesses ambiguity and the linguistic structure depends on many complex variables, including slang, regional dialects and social context.

Named entity recognition (NER) is the primary step towards information extraction in which an algorithm takes input as a string of text (sentence or paragraph) identifies relevant nouns (people, places, and organizations) that are mentioned in that string and classify named entities into pre-defined categories just like the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. NER is used in many fields in Natural Language Processing (NLP) and it can help in answering many real-world questions, such as:

- Which companies were mentioned within the news article?
- Were specified products mentioned in complaints or reviews?
- Does the tweet contain the name of a specific person?

Natural Language processing is considered to be a difficult problem in computer science. It's the nature of the human language that makes NLP complex to operate. Comprehensively understanding the human language requires understanding of both the words and how the concepts are connected to deliver the intended message. While humans can easily master a language, the ambiguity and imprecise characteristics of the natural languages are what make NLP difficult for machines to implement.

The paper presents implementation of NER and survey of various works in process of NER in the field of NLP. Related work and past literature is discussed in this section. The various techniques used in recognition of name entities, proposed techniques along with the challenges and the problems encountered are discussed in this paper.

II. LITERATURE SURVEY

The process of Named Entity Recognition consists of these stages: Stopword removal, Tokenization, Assign a tag to tokenized word, search for Ambiguous word and Entity Recognition. Disambiguation is completed by analyzing the linguistic feature of the word, its preceding word, its following word, etc. Considerable work is already done for foreign languages if we look at the same scenario for South-Asian languages such as Hindi and Marathi, it finds out that not much work has been done. As these languages are morphologically rich language and unavailability of annotated corpora.

In 2018, **Prince Rana, Sunil Kumar Gupta, Kamlesh Dutta** [1] proposed an approach for identifying the named entity where the processing unit is divided into two parts. First is the pre-processing task and second is the post processing task. Pre-processing of text includes tokenization of text followed by comparing the text with the online Hindi dictionary to check whether a token is a known or unknown word. If the token is an unknown word then post processing will perform action on the unknown word. The unknown word is compared with the implemented rules to check its identity otherwise the identity of unknown words is checked based on linked annotated corpus. They have achieved accuracy up to a certain level. The accuracy can be increased by increasing the size of the corpus and by handling the disambiguation.

Shrutika Kale and Sharvari Govilkar [2] proposed a survey on different techniques like Rule based Approach, Machine Learning Approach and Hybrid Approach form the major categories of the NLP NER algorithms. Out of the following categories it has been observed that the machine learning based approach is the best suited and most popular approach. In machine learning there are many sub categories of techniques such as HMM, CRF, SVM, ME. Based on different evaluation techniques and result analysis and also according to the review of their literature survey of experiments conducted across India by different researchers it had been proved that the Rule based, CRF, HMM are mostly implemented for Hindi, Marathi, Urdu, Punjabi, Bengali, Telugu. It has been observed that HMM, CRF gives the best results considering their limitations.

Shubhangi Rathod, Sharvari Govilkar [3] had presented a comparison of various POS Tagging techniques for Indian regional languages that had been done elaborately. They said that automatic POS tagging makes errors because many high frequency words of part-of-speech are ambiguous. Rule-based tagging assigns a word all possible tags and uses context rules to disambiguate statistical tagging assigns a word its most likely tag, based on the n-set values in a training corpus. Hybrid based tagging combines the two approaches.

Zhanming Jie, Aldrian Obaja Muis, Wei Lu [4] had used Dependency trees, which conveys crucial semantic-level information. In this work, they investigate how to better utilize the structured information conveyed by dependency trees to enhance the performance of NER. Unlike the present approaches which only exploits the dependency information for designing local features, that they had shown that certain global structured information of the dependency trees are often exploited while building NER models, where such information can provide guided learning and inference.Through extensive experiments, that they had shown that their proposed novel on dependency guided NER model performs competitively with models that supports conventional semi-Markov conditional random fields, while requiring significantly less period.

Simpal Jain and Nidhi Mishra [5] discusses a hybrid based approach for POS tagging on Hindi corpus. This paper is a review of different Techniques for Part of Speech tagging of Hindi language. The Hindi Word Net may be a rich resource, it's getting used by many Hindi Natural language processing (NLP) applications. Hindi Word Net consists of around 1 lakh unique class categories of words like Noun, verb, adjective, and adverb. But still, many words are not tagged, so they had used Rule based approach to assign tags to all words, and use context rules to disambiguate stochastic based approach, they had assigned the most likely tag to a word, based on the on-set values frequency in a corpus. Hybrid based tagging, which is a combination of the two approaches. They had concluded that Hybrid Approach provides higher accuracy, as compared to an individual rule based POS tagger and stochastic POS tagger.

Deepti Chopra, Nisheeth Joshi, Iti Mathur [6] have designed a NER system for the Hindi language using the Hidden Markov Model and got accuracy of 97%. They have explained the importance of HMM and its advantage. The accuracy is high but the size of the corpus is limited and the tagset is small. One more challenging aspect of HMM is it requires a lot of data for training.

Yavrajdeep Kaur, Er.Rishamjot Kaur [7] have designed a NER system by using Hybrid approach (combination of Rule Based Approach and List look Approach) for Hindi Language. The system is capable of extracting 10 named entities. Their accuracy is 95.77%. In this system they have added three new name entities that are money value, direction values and animal/bird entities. They concluded that by adding more entities and by increasing the size of the corpus accuracy can be increased.

Sachin Pawar, Nitin Ramrakhiyani, Girish K. Palshikar, Pushpak Bhattacharyy, and Swapnil Hingmire [8] over here they have used Distant Supervision framework, which is used to automatically create a large labeled data for training the sequence labeling model. The framework exploits a set of heuristic rules based on corpus statistics for the automatic labeling. Their approach puts together the benefits of heuristic rules, a large unlabeled corpus as well as supervised learning to model complex underlying characteristics of noun phrase occurrences. In comparison to simple English like chunking baseline and a publicly available Marathi Shallow Parser, their method demonstrates a better performance.

Amir Bashir Malik and Khushboo Bansal [9] have described the problems of NER in the context of Kashmiri Language and provide relevant solutions by using noun identification algorithm and named entity recognition algorithm. Not much research has been done for Kashmiri language, by more research on it can improve the accuracy.

Suvarna G Kanakaraddi, V Ramaswamy [11] they have used a new parser called fuzzy parsing which is less rigid than traditional parsing and appropriate for NLP. They have developed FLSR grammar for this parsing. The language used was C. This paper is made for English language. The input was English language; they generated permutations and on the basis of permutations the FSLR algorithm was applied. Fuzzy min max was to determine the degree of parsing. This parsing is made for partial sentences and gives partial syntactic correctness. The efficiency of the result or percent is not described.

Jenny Rose Finkel and Christopher D. Manning [12] have built a joint model of named entity recognition and parsing which is based on feature-based constituency parser. Their model produces a consistent output, where the named entity spans do not conflict with the phrasal spans of the parse tree. The joint representations allows the information from each type of annotation to improve performance on the other and they have done experiments on OntoNotes corpus and found

improvements of about 1.36% absolute F1 score for parsing, and up to 9.0% F1 score for named entity recognition. They said that they would like to add other levels of annotation available in the OntoNotes corpus to their model, including word sense disambiguation and semantic role labeling.

TABLE I : OBSERVATION

Sr. no	Language	Category	Accuracy
1	Hindi	12	97%
2	Hindi	Not Defined	95.77%
3	Punjabi	Not Defined	86.98%
4	Hindi, Marathi	Not Defined	66.05%
5	Bengali	Not Defined	53.36%
6	Kashmiri	Not Defined	93%

Above table shows the accuracies of different Indian languages for NER. Apart from [6], none of them have mentioned the tagset category and most of the papers are the survey paper and do not describe the techniques used for recognition of named entities. Deepti Chopra, Nisheeth Joshi, Iti Mathur [6] have used a HMM model for NER and have good accuracy but have a limitation of the dataset. Yavrajdeep Kaur, Er.Rishamjot Kaur [7] has used Hybrid approach (combination of Rule Based Approach and List look Approach) and has got good accuracy but has the same limitation of dataset for hindi.

III. EXISTING SYSTEM

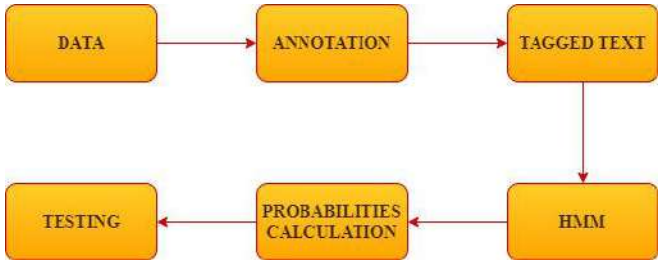


FIG 1. EXISTING SYSTEM ARCHITECTURE

In the above figure of the existing system architecture we can see the process in following steps:

1. **DATA:** The input data is provided to the system.

2. **ANNOTATION:** After preprocessing, annotation is done i.e. all tokens are identified. The name entity of the person, location, and organization and so on various annotations are done.

3. **TAGGED TEXT:** The annotated text is the tagged text which is used for training the model.

4. **HMM:** The tagged set is given to the Hidden Markov Model (HMM) for training.

5. **PROBABILITIES CALCULATION:** After training three probabilities are calculated (Start probability, Transition Probability and Emission probability) which are used to evaluate the model i.e. to know how good the model is.

6. **TESTING:** After training the model is tested to know how good the model is performing.

IV. PROPOSED SYSTEM

This system helps to find the NER i.e. the Named Entity Recognition of the input sentence. This helps to understand the relationship of the various terms in a sentence for example name, place, etc. This system is proposed for Indian regional languages i.e. Hindi, Marathi.

PROCESSING STEPS:

1. **INPUT:** The input data is provided to the system in the form of a document.

   Example:

   Input: भारतीय परंपरा के अनुसार आम आदमी आम खाता है |

FIG 2. NER USING SYNTACTIC PARSING FOR HINDI

2. **PRE-PROCESSING:** The input data is first processed by removing the stop words in it and carrying out tokenization, stemming and a root word is generated by the morphological analyzer.
   - **VALIDATION:** Validation is to check whether the given input text is in language for which the system is implemented. It also checks whether the input is syntactically correct, but does not check the semantic correctness.
     Comparing each input document character with UTF-8. If the character is present in UTF-8 then the input is a valid script(Hindi) for the language i.e. Hindi for which the system is implemented

Example:

Input: भारतीय परंपरा के अनुसार आम आदमी आम खाता है |

According to Indian tradition common man eats mango.(UTF-8,Font=Mangal)

Output: भारतीय परंपरा के अनुसार आम आदमी आम खाता है

- **STOPWORD REMOVAL:** In stop word removal, a word that occurs very frequently and does not contribute much to the context and content, and also have any impact on their existence are removed. Removing unnecessary words in the sentence such as से, को, इस, कि, जो, तो, ही, या, हो, etc. From a predefined list of stop words in the language.

Example:

Input: भारतीय परंपरा के अनुसार आम आदमी आम खाता है

Output: भारतीय परंपरा अनुसार आम आदमी आम खाता

- **TOKENIZATION**: The aim of the tokenization is the exploration of the words in a Sentence where every word, symbol, special character in the sentence is considered as a token.
  After removing stop words each character left is separated as a token.

Example:

Input: भारतीय परंपरा के अनुसार आम आदमी आम खाता है

Output: भारतीय | परंपरा | अनुसार | आम | आदमी | आम | खाता

- **STEMMING:** Trimming or cutting out the extraneous words to the stem is called stemming. Here inflections are removed using stemming algorithms. Here the suffixes and prefixes added to the root word are removed.

Example:

Input: भारतीय | परंपरा | अनुसार | आम | आदमी | आम | खाता

Output: भारती | परंपरा | अनुसार | आम | आदमी | आम | खाता

- **MORPHOLOGICAL ANALYSIS:** Morph analysis is the procedure to find out the root word. It recognizes the inner structure of the word.
  After stemming there is a possibility of not obtaining the exact root word in such cases the morphological analysis is important. The Morphological Analysis takes place with the predefined inflection rules in the system.

Example:

Input: भारतीय | परंपरा | अनुसार | आम | आदमी | आम | खाता

Output: भारत | परंपरा | अनुसार | आम | आदमी | आम | खाता

3. **POS TAGGING:** The parts of speech in the input data is identified and assigned to the word. It also removes the ambiguity in the sentence using word sense disambiguation.
   WSD will be used to remove the ambiguity in the sentences. Each token after stemming and morphing will be assigned with its POS in the sentence.

Example:

Input: भारत | परंपरा | अनुसार | आम | आदमी | आम | खाता

Output: भारत→NNP (Proper Noun)
परंपरा→ ABN (Abstract Noun)
अनुसार →CC (Conjunction)
आम→ JJ (Adjective)
आदमी→ NN (Common Noun)
आम→ NN (Common Noun)
खाता→ VM (Verb)

4. **SYNTACTIC PARSING:** Syntactic parsing is the task of recognizing a sentence and assigning a syntactic structure to it. In this case the input data will be parsed and a parse tree based on the Entity and relationships will be generated. A parse tree will be generated after POS tagging in which we can see the entity with its relation in the sentence.

Example:

<Sentence id="1">

1	((	NP	<fs af='परंपरा,n,f,sg,3,d,0_का_अनुसार vpos="vib2_3_4" head="परंपरा">
1.1	भारतीय	JJ	<fs af='भारतीय,adj,any,any,,any,,'>
1.2	परंपरा	NN	<fs af='परंपरा,n,f,sg,3,d,(name="परंपरा">
	))		
2	((	NP	<fs af='आदमी,n,m,sg,3,d,(head="आदमी">
2.1	आम	JJ	<fs af='आम,adj,any,any,,any,,'>
2.2	आदमी	NN	<fs af='आदमी,n,m,sg,3,d,(name="आदमी">
	))		
3	((	JJP	<fs af='आम,adj,any,any,,an head="आम">
3.1	आम	JJ	<fs af='आम,adj,any,any,,an name="आम">
	))		

4	((	VGF	<fs af='खा,v,m,sg,2,,ता_है,v vpos="tam1_2" head="खाता">
4.1	खाता	VM	<fs af='खा,v,m,sg,any,,ता,v name="खाता">
	))		

</Sentence>

5. **NER:** This parse tree constructed will be used to classify these entities which consist of proper nouns like person name, location names, temporal entities, etc. Clusters of different types of entities based on Name, Place, etc will be formed.

Example:

भारतीय परंपरा के अनुसार आम आदमी आम खाता है |
भारत → देश(Country)
आम → फल(Fruit)

V. TESTING AND ANALYSIS

In this system, software testing will be carried out on the basis of the end result of the system that is to obtain NER for Hindi language. The developed system classifies Entities based on Proper Nouns. The Entities include Name, Day, Date, Month, Country, States, Currency, Languages, Union Territory, States, Titles and City. The system only takes a .txt file as input. The testing is carried out with a group of sentences together in a document .The accuracy of the designed system is predicted and also understand the loopholes that are to be covered in the system to make it a very efficient system that can handle ambiguity of the language and complex nature of Hindi language.

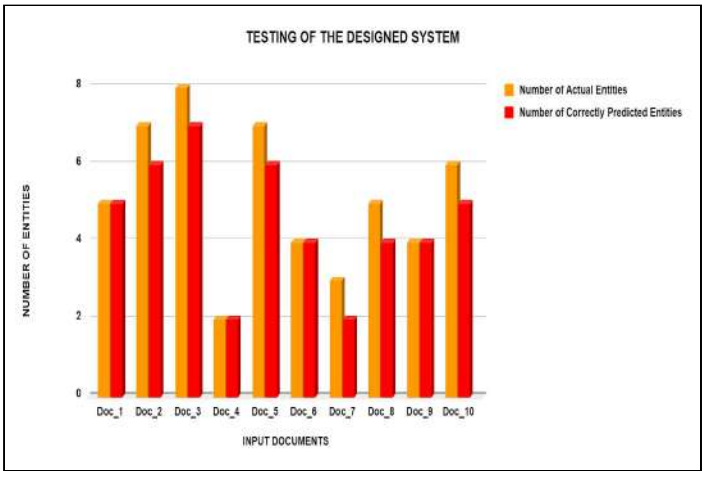


FIG 3. GRAPHICAL REPRESENTATION OF THE DESIGNED SYSTEM FOR DOCUMENTS

VI. RESULT

The below table II shows the accuracy obtained by the designed system for input of 10 different documents as input. The fig. 4 explains the accuracy obtained in the case.

TABLE II : ACCURACY OVERVIEW

For 10 Documents as Input	
Actual Entities	121
Predicted Entities	102
Non Predicted Entities	19
Accuracy(%)	84.31

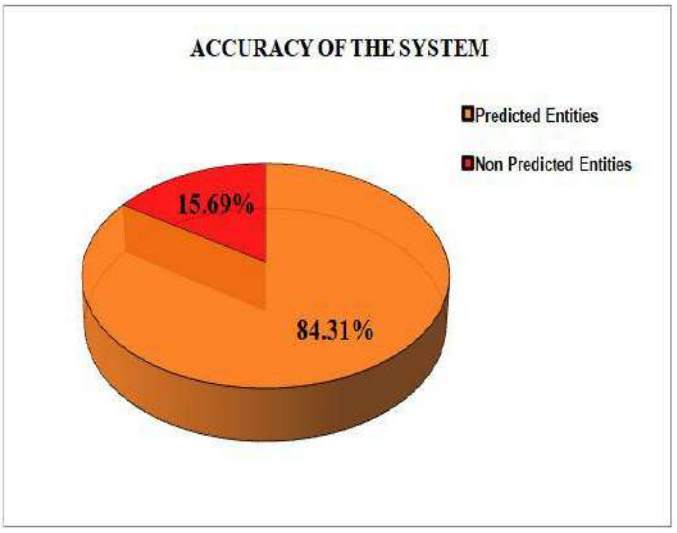


FIG 4. GRAPHICAL REPRESENTATION OF ACCURACY OBTAINED

In this system, after a lot of testing we find the overall accuracy of the system to be 84.31%. In the process it was seen that the system faces some difficulties in handling the ambiguity and nature of Hindi language. However, this can be removed with more testing and that would eventually increase the efficiency of the system. The system only takes .txt files as input and produces subsequent process outputs in the same manner. It should also be noted that very little work has been done on this field and can be used as a base for many other upgrades on future systems and can be used as a library package in the field of NLP.

VII. APPLICATIONS

Classifying content for news providers: A large amount of online content is generated by the news and publishing houses on a daily basis and managing them correctly can be a challenging task for the human workers. Named Entity Recognition can automatically scan entire articles and help in identifying and retrieving major people, organizations, places discussed in them. Thus articles are automatically categorized in defined hierarchies and the content is easily discovered.

Automatically Summarizing Resumes: You might have come across various tools that scan your resume and retrieve important information such as Name, Address, Qualification, etc from them. Also one of the challenging tasks faced by the HR Departments across companies is to evaluate a gigantic pile of resumes to shortlist candidates. A lot of these resumes are excessively populated in detail, of which, most of the information is irrelevant to the evaluator. Using the NER model, the relevant information to the evaluator can be easily retrieved from them thereby simplifying the effort required in shortlisting candidates among a pile of resumes.

Optimizing Search Engine Algorithms: When designing a search engine algorithm, It would be an inefficient and computational task to search for an entire query across the

millions of articles and websites online, an alternate way is to run a NER model on the articles once and store the entities associated with them permanently. Thus for a quick and efficient search, the key tags in the search query can be compared with the tags associated with the website articles

Powering Recommendation systems: NER can be used in developing algorithms for recommender systems that make suggestions based on our search history or on our present activity. This is achieved by extracting the entities associated with the content in our history or previous activity and comparing them with the label assigned to other unseen content. Thus we frequently see the content of our interest.

Simplifying Customer Support: Usually, a company gets tons of customer complaints and feedback on a daily basis, and going through each one of them and recognizing the concerned parties is not an easy task. Using NER we can recognize relevant entities in customer complaints and feedback such as product specs, department, company branch location so that the feedback is classified and forwarded to the appropriate department responsible for the identified product.

VIII. CONCLUSION

We have implemented the proposed system for the topic 'NER using syntactic parsing using NLP'. The whole project is created using python language and the GUI is using html and flask framework. The input is a text document and it is processed to find NER using Pre-Processing initially which consists of Validation, Stop Word Removal, Tokenization, Stemming, etc. carried by POS Tagging, Syntactic Parsing and NER. The study of different domain techniques is presented. The different techniques are studied and by studying them we came to know about the previous methods which were applied and the techniques used to carry out the name entity recognition for natural language processing. After testing the designed system we could obtain an accuracy of about 84.31%. However, we should also consider that based on more testing for ambiguity and developing new rules for this complex nature of the language the accuracy could be substantially increased in further upgradation. More ambiguity needs to be resolved with constant testing and penetrating to every complexity level of the language.

ACKNOWLEDGMENT

We would like to show our gratitude to each and every one part of the project and the project guide for sharing their proficiency with us during the course of this research, and we thank other faculty for their so-called insights. We are also immensely grateful to the authors of the referred references for their comments on an earlier version of the manuscript, although any errors are our own and should not tarnish the reputations of these esteemed persons. We would also like to thank the head of the Information Technology department and to the principal of Pillai College of Engineering, New Panvel for extending their support in this course of research.

REFERENCES

[1] Named Entity Recognition (NER) for Hindi Prince Rana, Sunil Kumar Gupta,Kamlesh Dutta International Journal of Computer Sciences and Engineering Vol.-6, Issue-7, July 2018 E-ISSN: 2347-2693

[2] Survey of Named Entity Recognition Techniques for various Indian Regional Languages Shrutika Kale and Sharvari Govilkar International Journal of Computer Applications (0975 – 8887) Volume 164 – No 4, April 2017

[3] Shubhangi Rathod et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (3) , 2015, 2525-2529 2017 Survey of various POS tagging techniques for Indian regional languages Shubhangi Rathod, Sharvari Govilkar

[4] Efficient Dependency Guided Named Entity Recogn -ition Zhanming Jie, Aldrian Obaja Muis, Wei Lu Singapore University of Technology and Design 2017

[5] Insight of various POS tagging techniques for Hindi Language Simpal Jain and Nidhi Mishra (Sept 2017)

[6] Named Entity Recognition in Hindi Using Hidden Markov Model Deepti Chopra , Nisheeth Joshi , Iti Mathur Second International Conference on Computational Intelligence & Communication Technology 2016

[7] Named Entity Recognition (NER) System for Hindi Language Using Combination of Rule Based Approach and List Look Up Approach Yavrajdeep Kaur , Er.Rishamjot Kaur International Journal of scientific research and management (IJSRM) 2015

[8] Noun Phrase Chunking for Marathi using Distant Supervision Sachin Pawar Nitin Ramrakhiyani Girish K. Palshikar Pushpak Bhattacharyya Swapnil Hingmire August 2015

[9] Named Entity Recognition for Kashmiri Language using Noun Identification and NER Identification Algorithm Amir Bashir Malik and Khushboo Bansal International Journal of Computer Sciences and Engineering Volume-3, Issue-9 E-ISSN: 2347-2693 2015

[10] Study of Named Entity Recognition for Indian Languages Hinal Shah, Prachi Bhandari, Krunal Mistry, Shivani Thakor, Mishika Patel and Kamini Ahir 2015

[11] Natural Language Parsing using Fuzzy Simple LR (FSLR) Parser Suvarna G Kanakaraddi, V Ramaswamy 2014 IEEE International Advance Computing Conference (IACC)

[12] Joint Parsing and Named Entity Recognition Jenny Rose Finkel and Christopher D. Manning Computer Science Department Stanford University 2009

Optimal safe route recommendation by examining roadside accident attributes

Ajay Dholapuriya, Abhishek Mallah, Yash Singh, Dr. Sushopti Gawade

Department of Information Technology, Pillai College of Engineering
New Panvel, Mumbai, MH, India

[1]Dholapuriyaajj17dse@student.mes.ac.in

[2]Mallahabr17dse@student.mes.ac.in

[3]yashjs15it@student.mes.ac.in

[4]sgawade@mes.ac.in

***Abstract*— with changing generation the thoroughfare issue is increasing greater in extent. Growing vehicular traffic causes road crashes, growth in the bad conditions of road in the developing countries. Thus safety on the wide way has become a vital issue as it directly affects human life. In our project, to tackle this problem, we will initially design a relative dataset. During pre-processing, the cleaned data will be transformed into an appropriate form for knowledge extraction, which in turn helps to find key attributes that lead to an accident on the roadside and perform analysis. Based on selected attributes, the accident-prone street will be detected by which our system will recommend the optimal safe route. The findings of this system can contribute information to the governing bodies for city street planning to reduce accidents.**

***Keywords*— Safe Route Recommendation, Accident Data analysis and Prevention, Data Mining, A* algorithm, Dijkstra's algorithm.**

I. Introduction

The World Health Organization reported, about 1.3 million people [10] die annually on the world's roads, and between 20 and 50 million non-fatal injuries. Pedestrians, cyclists, motorcyclists, etc. make up almost half of those killed on the roads. To prevent roadside accidents effectively accident-prone roads and risky time need to be determined. The analysis of vehicle accident spatio-temporal characteristics [1] might solve this problem.

The objective of this work is as follows; 1) The main objective is to reduce roadside accidents. 2) Analysing the spatio-temporal characteristics of the roadside. 3) Detecting accident-prone roads by using a danger index [1]. 4) Recommending the safe route to the user using a* algorithm. 5) It contributes to reducing traffic collisions.

After its successful implementation of this work, we able to provide these features: Safe route for users. Reduction in traffic collisions. Safe guarding travels from dangerous roads. It provides government and researchers with detailed visualization of motor collision in a concise manner. To provide mentioned features following tasks are completed.

1. Collecting required datasets.
2. Preparing Dataset for mining tasks.
3. Analysing the spatio-temporal characteristics of the roadside using K-Medoid algorithm.
4. Recommending a safe route.

As a first foot towards the research, we identified a city within the USA as an area of interest. The city is New York City. The reason to choose NYC as an area of interest because it has a large road network. And its major town within the US with associate calculable 19,979,477 folks in its Metropolitan applied math space and 22,679,948 residents in its Combined applied math space [10].

Thereafter, we took two datasets namely one was a collision dataset that was obtained from the New York police department official site [11]. Collision dataset had approx. 1.2 Million rows and had 29 columns. The second dataset was a spatial dataset gathered from the OpenStreetMap (OSM) site and had extension .osm. Data from OSM can be used in many ways such as creating an interactive map, routing and many more.

Rest of the paper is organized as follows: Section 2 describes the review of the relevant techniques. It describes the pros and cons of each technique. Section 3 presents the Implementation and proposed work. It describes the major approaches used in this work. The Conclusion and future scope of the report are presented in section 4.

II. Literature review

This section describes various methods through which we can deal with roadside accident problems. Each method has its

advantages and disadvantages. We will look at each of these solutions by dividing into two subcategories so that you can better understand each research methodology for road accident domain in less time.

A. *Data mining and Machine Learning*

Traffic accidents are a serious explanation for death globally, curtailing many lives per annum. Therefore, a system which will predict the occurrence of traffic accidents or accident-prone areas can potentially save lives. Salah Taamneh, Sharaf Alkheder and Madhar Taamneh [2] have design such system using artificial neural network. The objective of their study was used to predict the injury level of traffic accidents. They used the k-mean algorithm to improve the prediction accuracy of the ANN classifier. Significant improvement in the prediction has achieved after clustering. To compare the performance of the ANN model, an ordered probit model was used by [2]. Ordered probit model obtained the accuracy of 59.5% was less than the ANN accuracy. The only disadvantage of using this approach is that it cannot reveal spatio-temporal characteristics of accidents.

Data mining techniques such as classification, clustering may help us in characterisation of key factors, patterns related to traffic accident data were causally connected with different injury severity. Imad Mahgoub and Hamzah Al Najada [4] has applied Classification algorithms and feature selection techniques. The classifiers used in their research [4] are Naive Bayes, C4.5, Random Forest, AdaboostM1 (with the base classifierC4.5), and Bagging (with the base classifier C4.5). Out of which Naive Bayes gave optimum results. They also applied big data analytics to gain useful insights to increase road safety and decrease traffic crashes. In their paper [4], they used H2O and WEKA mining tools. By using a feature selection technique they found the most important predictors. They tackle the problem of class imbalance by employing bagging using different quality measures. Because of class imbalance, they obtained less accuracy.

In this study [3], numerous machine learning algorithms have been applied along with data mining for road accident estimation**.** The specific objective of their paper was to analysis for a status of the road accident occurrence and determination of the risk of an accident. They had used AdaBoost, CART, C4.5, Naive Bayes, OneR, IBk(Knn,) for estimates of road accidents. Kappa statistic, F-criterion, ROC Area were the performance measure to compare the accuracy of AdaBoost, CART, C4.5, Naive Bayes, OneR and IBk. According to their research CART, Ibk, C4.5, and Naïve Bayes algorithms were given a higher performance in Kappa statistic and F-criterion.

Exploratory visualization techniques and statistical analysis [5] sheds light on predicting the probability of accidents on roads with special emphasis on State Highways (SHs) and Ordinary district roads (ODRs) by estimating the severity of accidents supported the sort of accident. Pointing out the traffic collision data of roads the frequency of traffic collision of roads is analysed using correlation analysis and exploratory visualization techniques. Exploratory visualization techniques depict many crucial aspects such as frequency distribution of enormous data categories and summarize dataset in pictorial form. Limitation of their approach is that no analyse of a cause of the severity of accidents by considering other parameters such as non-restriction of speed, shoulder drop-off, etc.

Clustering and classification this technique is used by Authors Ayushi Jain, Garima Ahuja, Anuranjana and Deepti Mehrotra [7] to promote Road Safety in India. The objective of their paper is to have data mining to come to aid to create a model that not only smooths out the heterogeneity of the data by grouping similar objects together to find the accident-prone areas in the country with respect to different accident-factors but also helps determine the association between these factors and casualties. No determination of accident frequency.

This paper [8] proposed a safe driving suggestion to the road user using analysis of road traffic fatal accidents. For that, they did a careful analysis of road accident data which closely related to fatal accidents. They applied statistical analysis and data mining algorithms on the FARS Fatal Accident dataset to address this problem. To investigate the relationship between attributes like fatal rate collision manner, weather, surface condition, light condition, and drunk driver, association rule, classification model, K-means clustering algorithm wear used. Certain safety driving suggestions were made based on statistics, association rules, classification model, and clusters obtained.

B. *Optimal route*

Dijkstra's algorithm is the most popular pathfinding algorithm. Enbo Zhou, Mao and Shanjun Mao [1] had used K-medoid algorithms for investigating the spatio-temporal characteristics of motor vehicle collisions and Dijkstra for recommending the optimal safe route. Accident-prone streets were detected using dangerous index defined in this paper [1]. And to find different collision patterns they clustered the street based on the collision curve. The collision curve was obtained by clustering similar street based on accident frequency [1]. Bike accident has not been included and it's not generalized to all vehicular populations in their research. And they considered hours as time-related factors. The number of collisions varies day, week and month were not given in their study [1].

Traditional route recommendation systems have one main weakness. They usually recommend the same route for all users and cannot help control traffic jam. Which leads to more collisions. To address this problem, they [6] develop a route recommendation system using a* algorithm, a* algorithm utilizes historical taxi driving data and to provide users with different routes.

III. Implemented System

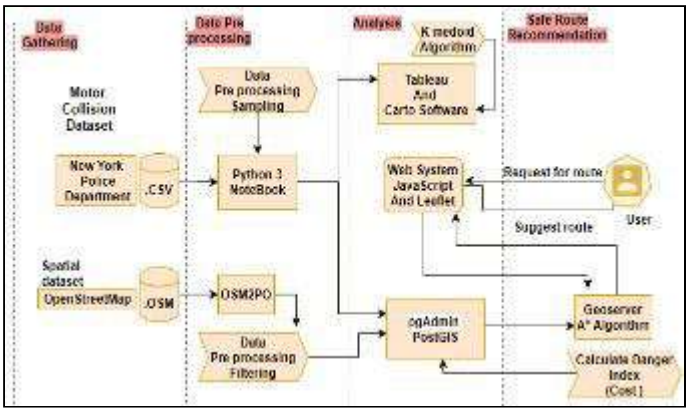


Fig. 1 Proposed system architecture

A. Data Gathering

As a first foot towards the initial step, we took two datasets;

1) Motor collision dataset of New York City (NYC) from the New York police department official site. This dataset gives an elaborate account of the road accidents that transpired in New York City. The dataset includes a total of 29 attributes, in which latitude and longitude, Street Name, Contributing factors types, Vehicle type played an important role.

2) Spatial dataset of NYC gathered from the OpenStreetMap site and has an extension .osm. This dataset was used for routing. The result of this section was two datasets. Each dataset was given to data pre-processing block

B. Data Pre-processing

Data pre-processing is a phase in which unwanted, redundant, missing data's are removed from old dataset [4]. Then algorithms like clustering, regression, pattern recognition etc. are performed on new dataset to derive insights.

The same phase is carried on our two datasets using two different tools.

1) Motor collision dataset had approx. 1.2 Million rows and had 29 columns before pre-processing. 289k rows and 20 columns after pre-processing. This task was performed using Google Colab Environment of python 3. This dataset has some important attributes contributing factors for accidents like driver inattention, Vehicle type, Time and date and Latitude and Longitude. After pre-processing it's passed to Tableau and Carto for further tasks.

2) A Spatial dataset cannot be directly used for routing applications. So we again did the pre-processing of spatial dataset data using the Osm2Po tool. We pre-processed data so that osm data becomes routable and has a street name and its coordinates. After pre-processing it we passed it to the postGIS database. Where testing was done to ensure it's routable.

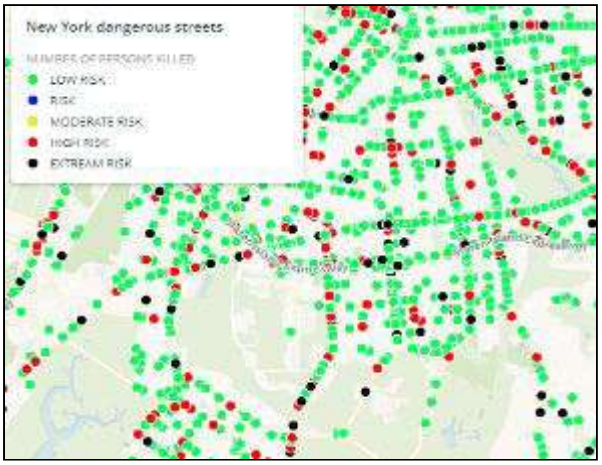


Fig. 2 clustering result

C. Data analysis

We used the K-medoid Clustering to group similar roads. As per this algorithm, a street can be allotted to only one cluster. In K-medoid K is the number of clusters and is usually given a small integer value (1, 2, 3...). K points are then chosen randomly-preferably the initial ones which represent the centroids of k clusters without any members. Fig 2 shows the result of clustering.

Some Observations after clustering are:

Weekday analysis:

Most of the accidents happen on weekdays. Friday looks a little peak day with approx. 28k accidents. Weekends are safe to drive.

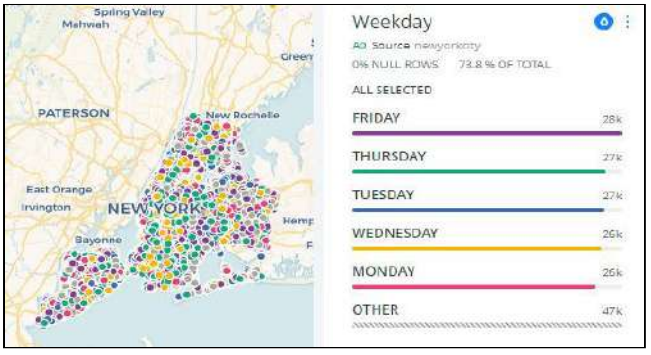


Fig 3. Week day

Month analysis:

June, May and July reported the equal but highest number of accidents with approx. 18k. These months are summer season in New York. Hence summer is more danger that winter and monsoon in New York City.

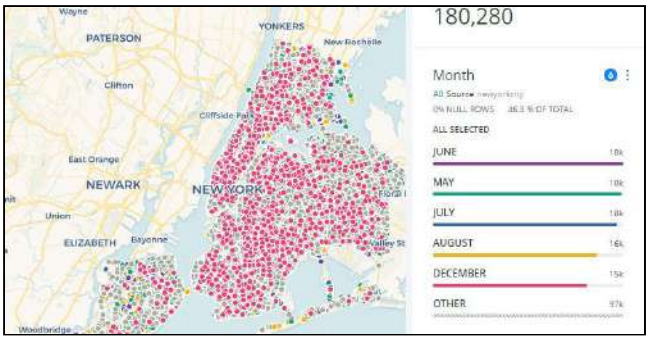


Fig 4. Month day

Key factors:

The two most common top causes of accidents are driver distraction and failure to yield right of way.

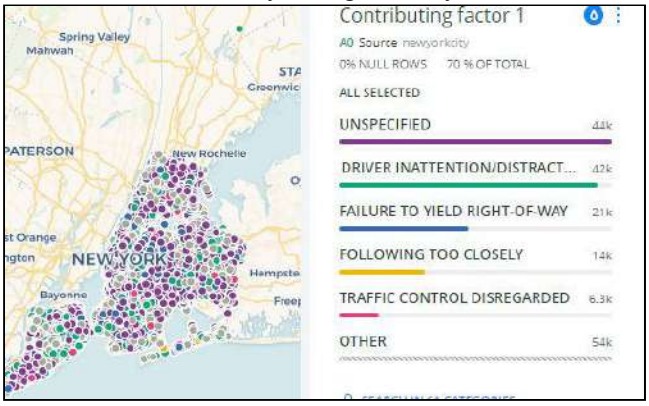


Fig 6. Key factor

Quarter analysis:

Same story here also saying second quarter has a high collision rate than other. Second-quarter of every year from 2012 to 2019

That nothing but the summer season.

Reason for summer to have more may be due to large of trave -llers coming to or going from New York City.

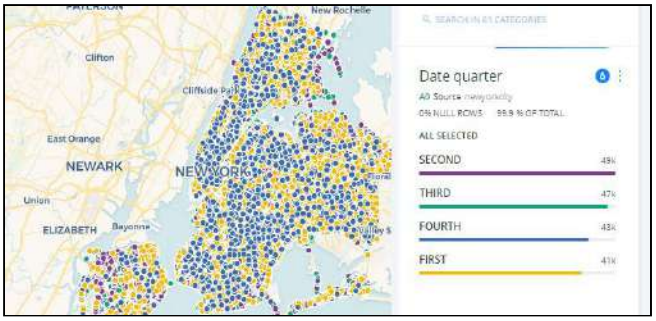


Fig 5. Date quarter

Vehicle type:

The two most common types of vehicles involved in a collision are sport utility/station wagon and a passenger vehicle.

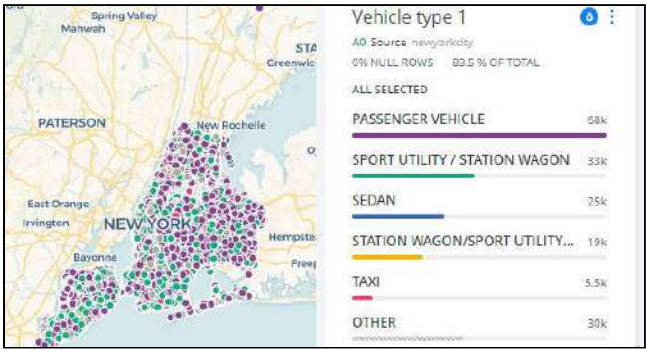


Fig 7. Vehicle Type

A better interactive data visualization all result provided here [13].

D. Safe Route Recommendation

Both datasets were coupled using postGIS.

Fig 8. User Interaction with System

To find accident-prone street Danger Index was used, which is calculated by formula:

D.I = Collision Number /Street Length

Collision number reflects number of motor collision on each street divided by its street length. Street length is expressed in km. After calculating D.I for each street we passed it to geoserver where it acted as a cost for safe route algorithm. We selected a* algorithm as safe route algorithm. The reason to choose a* algorithm over Dijkstra's algorithm is that: a* always tries to improve runtime to find an optimal solution. Fig 3 shows the user interaction with the system. User will ask for the safe route. If a route exists geoserver return on Web interface or else it displays not found. Geoserver will return two routes one is the safe route and another is normal.

IV. Conclusion and Future scope

A. Conclusion

1) Our finding includes two aspects mainly as follows firstly many collisions occur on weekdays, summer, and in a second quarter of year. Friday interestingly is a little peak.

2) This paper's methods and findings may contribute optimal safe routing method provides a replacement insight to route

under some circumstances like school bus's travel. Furthermore, the numbers of collision's variations by day, week and month are given.

3) The spatio-temporal scale can be extended which may provide more information to the government.

B. Future work

1) The route finding interface generates routes after dragging source and destination markers, we can further add the edit boxes where users can be select source and destination after that marker and route will be seen on a map.

2) Real-time motor collision dataset can used which provide up to date collision information.

ACKNOWLEDGMENT

We are very thankful to our project guide Dr. Sushopti Gawade, for her valuable support, constant guidance. Regardless of her extremely busy schedule, she never failed to take time out for us to help solve our problems and clear our doubts. We would like to take this opportunity to thank our Head of Department, Dr. Sastishkumar L. Verma for his heartening, motivation, guidance and support. We would also like to thank Dr. Sandeep M. Joshi, Principal, PCE, New Panvel for his invaluable support and for providing an outstanding academic environment. We acknowledge all the faculties of the Information Technology Department for their advice during various phases of this project work.

REFERENCES

[1] Enbo Zhou, Mao and Shanjun Mao, "Investigating Street Accident Characteristics and optimal safe route recommendation: a case study of New York City", 25th International Conference on Geoinformatics, Page(s):1- 7, 2017.

[2] Sharaf Alkheder, Madhar Taamneh and Salah Taamneh, "Severity Prediction of Traffic Accident Using an Artificial Neural Networks", Journal of Forecasting, J. Forecast. 36, 100–108 (2017).

[3] Halil İbrahim BÜLBÜL, Tarık Kaya, Yusuf Tulgar "Analysis for Status of the Road Accident Occurrence and Determination of the Risk of Accident by Machine Learning in Istanbul", December 2016.

[4] Hamzah Al Najada, Imad Mahgoub, "Big Vehicular Traffic Data Mining: Towards Accident and Congestion Prevention", 2016 International Wireless Communications and Mobile Computing Conference (IWCMC).

[5] Ms.Gagandeep Kaur, Er. Harpreet Kaur "Prediction of the cause of the accident and accident-prone location on roads using data mining Techniques", 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT).

[6] Henan Wang, Guoliang Li, Huiqi Hu, Shuo Chen†, Bingwen Shen† Hao Wu‡, Wen-Syan Li, Kian-Lee Tan, "R3: A Real-Time Route Recommendation System", Proceedings of the VLDB EndowmentAugust 2014 https://doi.org/10.14778/2733004.2733027.

[7] Ayushi Jain, Garima Ahuja, Anuranjana and Deepti Mehrotra "Data Mining Approach to Analyse the Road Accidents in India", 2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA).

[8] Liling Li, Sharad Shrestha, Gongzhu Hu "Analysis of Road Traffic Fatal Accidents Using Data Mining Techniques", IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA), 03 July 2017.

[9] "New York City", Wikipedia, 29-1-2020, [Online]. Available: https://en.wikipedia.org/wiki/New_York_City. [Accessed: 29-2-20].

[10] "Violence and Injury Prevention", World Health Organization, available:https://www.who.int/violence_injury_prevention/publications/road_traff. [Accessed: 29-2-20].

[11] "Motor Vehicle Collisions crash", NYC open data, 20-9-2019, [Online].https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95/ [Accessed: 20-9-2019].

[12] "New York City Map", Ajay Dholpuriya, 20-1-2020, [Online].Available: https://ajayanay.carto.com/builder/9dd4ca99-e949-4808-b648ae409f861a0b/widget/7a05fde1-f587-4db2-8190-08004d611d62 [Accessed: 16-2-20].

OBJECT DETECTION AND IDENTIFICATION (TRAFFIC SIGNS AND SIGNALS)

Gaurav Nikam*, Krutika Parvatikar , Neha Patil****

* Information Technology, Pillai College Of Engineering

Abstract Deep Convolutional Neural Network (CNN) have shown impressive performance in various vision tasks such as image classification and object detection using object detection APIs. For object detection, particularly in still images, the performance has significantly increased. In this work, we will be introducing a complete framework for object detection task in video domain (VID), in which object location at each frame is required to be annotated with the next line.. A complete framework is implemented for the VID task based on still-image object detection and general object tracking. In this video domain (VID), a stack of images is concatenated on top of each other and this stack of images fromw7 the said video given as input to the CNN. Their performance easily stagnates by constructing complex ensembles which combine multiple low-level image features with high-level context from object detectors and scene classifiers.

I. INTRODUCTION

As Artificial Intelligence technology advances, the application of automated driving technology in vehicles has attracted massive attention. Traffic sign and signal recognition system is a vital subsystem of the automated driving system. And Traffic sign and signal detection is the key technology of the traffic sign recognition system. In this project we will be using CNN based concept which is one of the main categories to do image recognition, image classifications. Objects detection, recognition faces etc. are some of the areas where CNNs are widely used. Basic procedure is that a video will be captured and then CNN will classify the objects present in that video in the form of pixels and frames. CNN image classifications takes an input image, process it and classify it under certain categories. Computers sees an input image as an array of pixels and it depends on the image resolution. Based on the image resolution, it will see h x w x d(h = Height, w = Width, d= Dimension).This paper proposes a method based on the Faster R-CNN deep learning framework to use traffic sign and signal detection.

II. IDENTIFY, RESEARCH AND COLLECT IDEA

We reviewed some implementation techniques like EasyNet Model, Bounding Box method, Unified method. After thoroughly reviewing and understanding these methods we came to the conclusion that Hybrid approach using Bounding box and Unified method is the best approach for our project. We physically collected datasets of traffic signs and signals by clicking real-time photographs and taking videos as our project is based on video, which is processed as images in frames for output. We referred to the German dataset(GTSDB) to compile our dataset. We used Labelimg tool to label our dataset and formed separate datasets for Traffic Signs and Signals.

III. METHODOLOGY

Proposed system architecture:

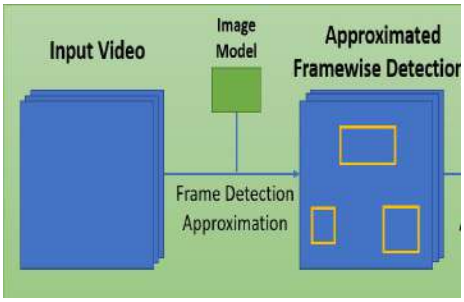
A Convolutional neural network(CNN) is a deep learning algorithm which can take in an input images, assign importance(learnable weights and biases) to various aspects/objects in the image and be able to distinguish one from the other. The preprocessing required in a ConvNets is much less as compared to other classification algorithms. While in primitive methods filters are hand-engineered, with enough training, ConvNets have ability to learn process these filters/characteristics.

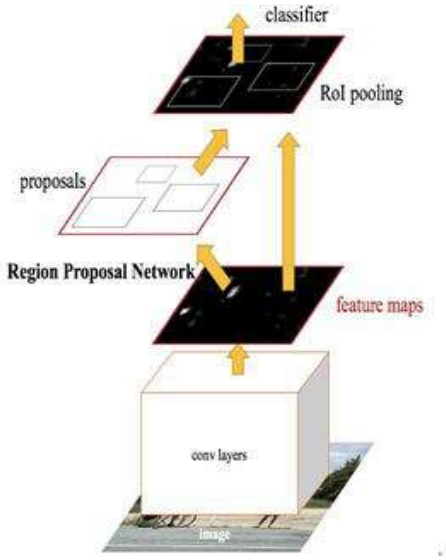
CNN algorithm is preferred specifically for image classification that is it will only classify the image and not the region the image primarily. To identify the detected region we require R CNN algorithm. The difference between object detection algorithms and classification algorithms is that in detection algorithms, we try to draw a bounding box around the object of interest to locate it within the image and also identify its region.

R-CNN

To bypass the problem of selecting a huge number of regions there exists a method where we use selective search to extract just 2000 regions from the images, called as region proposals. Therefore, now, instead of trying to classify a huge number of regions, we can just work with 2000 regions. These 2000 region proposals are generated using the selective search algorithm. There is one more better approach for object detection which is called as Fast CNN. The reason “Fast R-CNN” is faster than R-CNN is because you don’t have to feed 2000 region proposals to the convolutional neural network every time. Instead, the convolution operation is done only once per image and a feature map is generated from it which enhances the process.

Implementation details:





Step 1: Gather Pictures
TensorFlow needs tons of images of a particular object to train an efficient detection classifier. The training images should have random objects in the image along with the desired objects, and should have diverse backgrounds and different natural light conditions. There should also be some images where the desired object is partially obscured, halfway in the picture or overlapped with something else.

Step 2: Preprocessing of Data
In this system, the datasets will be of different traffic signs and traffic signals. Then, we pre-processed the data pertaining to the proposed project. These pre-processed images were used for detecting the same using object detection algorithms. Then, the images were trained and tested accordingly.

Step 3 : Label Pictures
With all the pictures gathered, we used Labelimg tool to label the desired frames in every image of the dataset.

Step4: Generating Training Data
After Labellimg the images, we generated the TFRecords that serve as input data to the TensorFlow training model.
First, the image .xml data was used to create .csv files containing all the data for the train and test images. Commands were issued on the editor to run the training.

Step5: Run the training
From the \object_detection directory, we issued the command to begin training. After everything has been set up correctly, TensorFlow initialized the training. The initialization can take up to 30 seconds before the actual training begins.

Step6 : Accessing the inference graph
Now that training is complete, the last step we performed was to generate the frozen inference graph (.pb file). From the \object_detection folder where it should be replaced with the highest-numbered .ckpt file in the training folder. This creates a frozen_inference_graph.pb file in the \object_detection\inference_graph folder. The .pb file contains the object detection classifier.
This classifier is our implemented classifier for object detection. This classifier can now be used to test out an image or a video using a webcam.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

The algorithm in this article trained the Faster R-CNN model using physically collected datasets. The data set has 1,706 images, including 1,478 training images and 228 test images.The signs in the image include various categories: parking, no parking, red signal, yellow signal and green signal . In this paper, the experiment using the Python language version of the TensorFlow deep learning system in Intel ® core i5 7th Gen (R) RADEON ® Windows 10 system configuration. We adopted both the average precision (AP) and the mean average precision (mAP) as the evaluation indexes. The AP is the area under PR (Precision-Recall) curve and the mAP is the mean of the AP for all categories. The experimental results which we got is 60-70% that indicates the method in the paper achieves a good performance.

V. FUTURE SCOPE

It should be noted that object detection has not been used much in many areas where it could be of g reat help. As mobile robots,and in general autonomous machines, are Authors

starting to be more widely deployed (e.g., quad-copters, drones and soon service robots), the need of object detection systems is gaining more importance. Finally, we need to consider that we will need object detection systems for nano-robots or for robots that will explore areas that have not been seen by humans, such as depth parts of the sea or other planets, and the detection systems will have to learn to new object classes as they are encountered. In such cases, a real-time open-world learning ability will be critical.

VI. CONCLUSION

The traffic sign and signal detection algorithm proposed in this paper makes use of faster r-cnn deep learning based technology to automatically extract features and realizes end-to-end training mode by using RPN. The algorithm in this article is carried out on physically collected data set, and has achieved a good detection effect.

ACKNOWLEDGMENT

We would like to thank our mentor Prof.Varunakshi Bhojane for her guidance and unwavering support throughout the project and the semester.
We would like to thank our HoD, Dr.Satishkumar Varma for his encouragement and motivation to learn and implement projects of sorts.

Lastly, we would like to thank our principal, Dr.Sandeep Joshi for providing us opportunities to explore our domain and for motivating us to do better.

REFERENCES

[1] Pratik Kalshetti, Ashish Jaiswal (Indian Institute of Technology, Bombay),"Object Detection Project Report"(CSE IIT-B 2018)

[2] Ross Girshick,Jeff Donahue, Jitendra Malik "Rich feature hierarchies for accurate object detection" (CVPR 2016)

[3] "You only look once: Unified real time object detection" Redmond Joseph, Computer Vision and Pattern Recognition, 2016

[4] Sandeep Kumar,Aman Balyan,Manvi Chawla (2017). Object Detection & Recognition in Images (Indian Institute of Technology,Delhi) IJEDR,2018

[5] Stan Lee(2018). Object Detection of Pinochle Deck of cards using TensorFlow API

[6] H. Bay, T. Tuytelaars, and L. van Gool, "Surf: Speeded up robustfeatures," Computer Vision and Image Understanding, vol. 110, no. 3,pp. 346–359, 2012

[7]H. Ishida, T. Takahashi, I. Ide, Y. Mekada, and H. Murase, "Identificationof degraded traffic sign symbols by a generative learning method," inProc. 18th Int. Conf. Pattern Recognition (ICPR 2006), vol. 1, 2006, pp.531–534.

[8] A. Broggi, P. Cerri, P. Medici, P. P. Porta, and G. Ghisio, "Real time roadsigns recognition," in Proc. IEEE Intelligent Vehicles 2007 Symposium,2007

First Author – Gaurav Nikam, BE IT, Pillai College Of Engineering, email id – nikamgashit16e@student.mes.ac.in.
Second Author – Krutika Parvatikar, BE IT, Pillai College Of Engineering, email id – parvatikarksit16e@student.mes.ac.in.
Third Author – Neha Patil, BE IT, Pillai College Of Engineering, email id – patilnmit16e@student.mes.ac.in.

Detecting Key-Needs in Crisis

Sagar Kulkarni
Department of Computer Engineering,
Pillai College of Engineering,
New Panvel
(*University Of Mumbai*)
skulkarni@mes.ac.in

Shailesh Gupta
Department of Information Technology,
Pillai College of Engineering,
New Panvel
(*University Of Mumbai*)
guptashhait16e@student.mes.ac.in

Abhay Gupta
Department of Information Technology,
Technology
Pillai College of Engineering,
New Panvel
(*University Of Mumbai*)
guptaabmeit16e@student.mes.ac.in

Navin Joshi
Department of Information
Pillai College of Engineering,
New Panvel
(*University Of Mumbai*)
joshinnit16e@student.mes.ac.in

***Abstract*—When a crisis occurs, the world springs into action to try and understand what is happening and what help is required. During these times, social media has become a key avenue through which to disseminate information. Twitter is one such social media having users across the world used during such times where the information is shared in form of texts called "tweets". But along with tweets containing vital information comes in other non useful tweets like sympathy tweets, political blaming, etc. As fast as this information is processed to find the needs of the victims, more quick authorities and those who want to help can act. During such times, the system will classify Tweets into different categories such as volunteering,disease transmission, donations ,injury ,death count, sympathy, irrelevant etc. The system uses ML models trained on large datasets of tweets related to disasters. It is trained using the SVM algorithm. This projected is currently confined to tweets in English language. The system accepts a hashtag as a input to fetch real time tweets using Twitter API to do real-time classification. The tweets are segregated into different CSV files packed inside a ZIP file for users to download.**

Keywords—NLP, ML, Tweets, Tweet analysis, Disasters, Natural Calamities, Disaster Management, Disaster relief, Information extraction, Information Segregation, Real-time tweet analysis.

I. INTRODUCTION

Whenever some disaster occurs it's information needs to be distributed to the rest of the world to get help. Off late in such conditions whenever most of the communication media fails the internet remains to be a last resort to spread the information about the situations. People use social media channels to share updates regarding the event. But since this data keeps flowing a system is needed to extract information from it to make more use of data and make help available as quickly as possible. All the updates need to be classified into different types to fasten the process of help. An automated system which can classify the data of disasters into different categories in real time as data comes in without manual help needs to be developed.

A.Objectives

a) To Extract an event or key element of Crisis such as Earthquake, Tsunami, Political issue, Terrorist attack etc from the given stream of data and provide

the response requirement to that Crisis .
b) Collecting the data from different resources such as news and media platforms.
c) Deciding which machine learning algorithm will give the best result and accuracy.
d) The Objective is to find key elements in different languages and provide automated machine response by using some Machine Learning Technique.
e) To drastically decrease the time and increase the efficiency while doing so.

B.Scope

The scope of this project is to provide help to the disaster occur Location population. The help provides such as a number of emergency contacts, help from different private NGO and Government NGOs. The System can be deployed for security purposes in police stations to identify the location where the disaster has occured. The application can also be used to analyse how much damage has occurred in that particular location .

II. RELATED WORK

Event extraction from tweets and social media are done previously but they have a few limitations. In this section we will review the previous system and approaches to find their strengths and their limitations; Moumita Basu, Saptarshi Ghosh and Kripabandhu Ghosh suggested a system which is used to identify the disaster occurring at some places using different social media . They are also checking for the fact that news will be fake or genuine. They are using the binary class classification for identifying the fact message or non-fact message. [2]
Observation: The dataset will be bigger to train our model and also it will not recognize local language. Location of the disaster position needs to be extracted.
Tulsee Doshi, Emma Marriot and Jay Patel developed a model to classify tweets into key categories like volunteer services, Displaced people and activations etc. The dataset consists of tweet IDs from 19 disasters representing 8 types of crisis in Spanish and English.After preprocessing the common twitter slang was replaced with complete terminology.A model based on Feed Forward Classifier on single tweets outperformed LSTM on single tweet as well as sequence of tweets. The accuracy increased with Word2Vec vectors instead of predefined vector set like Glove. [7]
Observation: Small dataset creates an overfitting problem because the dataset being small the model cannot learn efficiently about different features. So it fails on any data other than the testing data. Time series tweets accuracy is not good.

III. SYSTEM ARCHITECTURE

A. Overview

The Section presents an Overview of techniques used for the system. In this project the data will be collected from Twitter API or different social media platforms using web scraping. After collecting the data it will be processed to convert into suitable form so that model can be trained on it. The model will be trained using different disaster data to classify different categories. After building the model, test data will be given to the model to test it's accuracy. Model performance and accuracy will also be visualized.

B. Existing System

Existing solutions only classify tweets whether they are related to a disaster crisis or not. In Existing Architecture, they are not using as much as large amounts of data so that accuracy or predictability of will be less. [1]

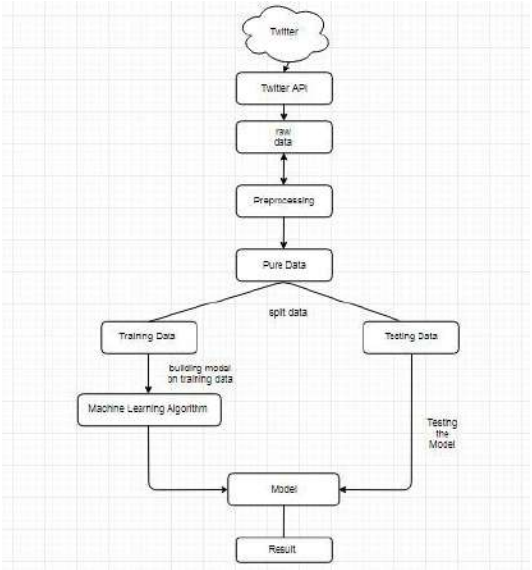


Figure 1 Existing System Architecture

C. Proposed System

In our proposed system we try to extract key needs from the classified tweets. We will try different algorithms and find their accuracy and performance. The model which gives the best result will be implemented.

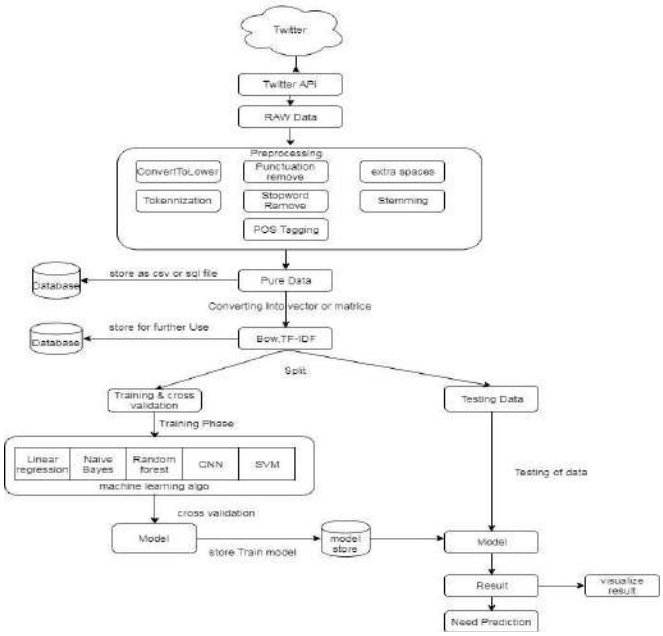


Figure 2 Proposed System Architecture

D. Sequence of steps involved

1) Collection of data:

a) Web Scraping: Web scraping is a term for various methods used to collect information from across the Internet. Generally, this is done with software that simulates human Web surfing to collect specified bits of information from different websites.

b) Twitter API: Twitter API is a RESTful service provided by Twitter to post and retrieve tweets. Twitter's standard search API (search/tweets) allows simple queries against the indices of recent or popular Tweets and behaves similarly to, but not exactly like the Search UI feature available in Twitter mobile or web clients.

2) Cleaning the data:

Tweets are terse and noisy in nature. Cleaning is done to remove all the noisy data from the dataset and to make the data suitable to feed to machine learning algorithms. This involves steps like:

a)Tokenization: Tokenization is the process of splitting the given text into smaller pieces called tokens. Words, numbers, punctuation marks, and others can be considered as tokens.

b)Convert to lower case: It is used to convert all the letters into Lowercase. Generally computers treat upper and lower case of the same letter as different. Therefore if all the text is converted to lowercase the text matching improves.

c)Punctuation Removal : The punctuation marks are removed from the text because they add no meaning to the data thus of no use.

d)Blank/White space Removal :To remove leading and ending spaces, you can use the *strip()* function. It is helping to reduce the memory uses and increase the efficiency of the model.

e)POS Tagging :Part-of-speech tagging aims to assign parts of speech to each word of a given text (such as nouns, verbs, adjectives, others) based on its definition and its context.Part-of-speech tagging is also referred to as word category disambiguation or grammatical tagging. PoS tagging is used in natural language processing (NLP) and natural language understanding (NLU).

f)Stemming/Lemmatization: The aim of lemmatization, like stemming, is to reduce inflectional forms to a common base form. As opposed to stemming, lemmatization does not simply chop off inflections. Instead it uses lexical knowledge bases to get the correct base forms of words.

g)Stop-Word Removal: The removal of the Stopwords (such as “is”,”the”.etc) is called Stop Word Removal. Stopwords add very little meaning so if removed the database space is saved and processing speed improves.

h)Chunking:Chunking is a natural language process that identifies constituent parts of sentences (nouns, verbs, adjectives, etc.) and links them to higher order units that have discrete grammatical meanings (noun groups or phrases , verb groups , etc.)To identify the different part of sentences and to find out the similarity and meaning of the sentence.

3) Convert text into vector form:

Text data cannot be operated on by machine learning models. Therefore, we need to convert the text data into numerical form before feeding it to the machine learning models. We convert the text data into vectors using Word2Vec, TF-IDF and BoW algorithms.

4) Train test splits:

Now that we have converted the text data to vector form, we split the data into train data and test data in the ratio 9:1. So 90% of data will be used for training and the remaining 10% will be used for validation of the model and getting the model's accuracy.

5) Training the model:

Using various machine learning classification algorithms we train the training data. Algorithms used are:

a)Naive Bayes: The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used Because of often sophisticated classification work. Using Bayes theorem we can find the probability of c (class) given that x has already occurred.

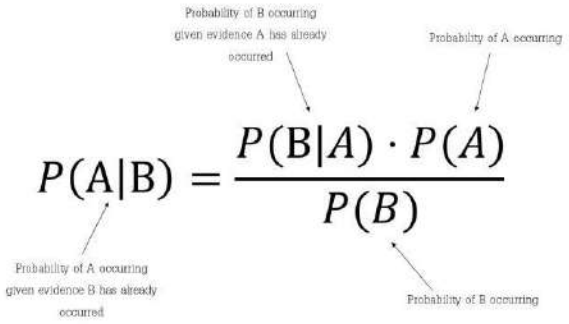


Figure 3 Bayes Theorem

b)Logistic Regression: Logistic Regression (LR) is a Generalized Linear Model (GLM). Although in spite of its name, the model is used for classification, not for regression. Logistic Regression is a probabilistic algorithm.

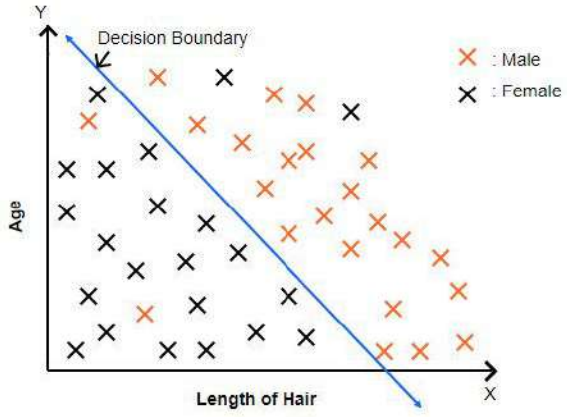


Figure 4 Example of Logistic Regression

c)SVM: A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. Here two support hyperplanes are drawn at extreme points to draw a margin between two classes. This helps increase the marginal accuracy and memory efficiency.

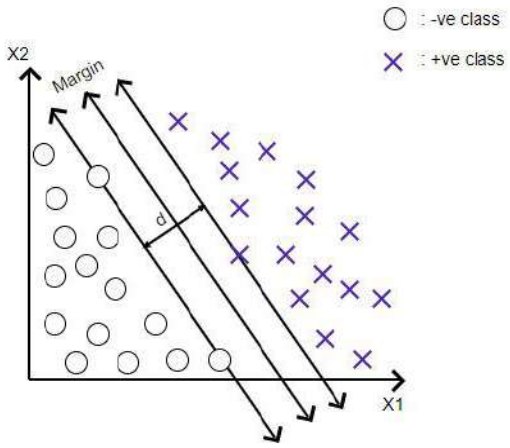


Figure 5 Example of SVM

d)Random Forest: Decision Trees are a class of very powerful Machine Learning model cable of achieving high accuracy in many tasks while being highly interpretable. What makes decision trees special in the realm of ML models is really their clarity of information representation. The "knowledge" learned by a decision tree through

training is directly formulated into a hierarchical structure.

6) Prediction on test data:
After training, we predict the label for the test data and compare it with the actual data to evaluate the model and find its accuracy using accuracy metrics.

Sr No.	Machine Learning Algorithm	Accuracy
1	Multinomial	56.32
2	Bernoulli Naive Bayes	59.33
3	Logistic Regression	73.00
4	SGD	74.93
5	SVC	53.47
6	Linear SVC	75.47
7	Random Forest	33.86

Figure 6 Accuracies of models

7) Results:
We notice, Linear SVM gives the most accurate results. So we now save our model in a pickle file so we don't have to build it again.

E. Hardware and Software Specifications
The experiment setup is carried out on a computer system which has the different hardware and software specifications as given in Table 1 and Table 2, respectively.

Table I. Hardware details

Type	Minimum Requirement
Processor	1.8Ghz Intel/Amd
RAM	4GB
Graphics	512MB

Table II. Software details

Type	Minimum Requirement
Programming Language	Python 3.6, Html, Css
IDE	IDLE/jupyter
Operating system	Windows 7 and up.
Database	Mysql
Browser	Chrome

IV. PROJECT INPUTS AND OUTPUTS

A. Input Details
The project takes hashtags as input. This input is processed through the back-end for display of results.

Figure 7 Input Interface

B. Output Details

Figure 8 Initial Output Interface

Figure 5 shows the output of the initial screen when a hashtag is typed as an input.

Figure 9 Final Output Interface

Figure 6 shows the output when live tweets are classified under different labels.

V. SUMMARY

In this paper, study of different Natural Language Processing techniques is presented. Different preprocessing techniques such as tokenization, upper to lower case, punctuation removal, blank/white space removal, POS tagging, stemming/lemmatization, stop word removal and chunking are explained.

Studies of various machine learning algorithms such as Naive Bayes, Logistic Regression, SVM and Random Forest are presented. A comparison of these algorithms based on their accuracies on our data is also depicted.

Applications of this domain are identified and limitations of previous systems are rectified.

VI. FUTURE SCOPE

To improve the accuracy of the model. To develop an interface to allow direct donations. To identify key elements in different languages.

REFERENCES

[1] Pattabhi RK Rao,Sobha Lalitha Devi, "Event Extraction from Newswires and Social Media Text in Indian Languages", FIRE2018.

[2] Moumita Basu, Saptarshi Ghosh,Kripabandhu Ghosh, "Information retrieval for microblogging during disaster" , IRMiDis FIRE2018.

[3] Chintak Mandalia,Memon Mohammed Rahil,Manthan Raval,Sandip Modha, "Entity extraction from social media indian language" , FIRE2018.

[4] Banujan. K1, Banage T. G. S. Kumara, Incheon Paik, "Social Media Mining for Post-Disaster Management – A case study on Twitter and News", International Research Conference on Smart Computing and Systems Engineering - 2018.

[5] Satya Katragadda , Ryan Benton , Vijay Raghavan, "Sub-Event Detection from Tweets", International Joint Conference on Neural Networks (IJCNN) 2017.

[6] Arpan Mandal, Kripabandhu Ghosh, Arnab Bhattacharya, Arindam Pal, Saptarshi Ghosh, "Information Retrieval from Legal Documents (IRLeD)", FIRE-2017-IRLeD.

[7] Tulsee Doshi, Emma Marriott, Jay Patel, "Detecting Key Needs in Crisis" , Stanford Education 2017.

[8] Peter Bourgonje, Julian Moreno Schneider, Georg Rehm, "From Clickbait to Fake News Detection: An Approach based on Detecting the Stance of Headlines to Articles", DFKI GmbH, Language Technology Lab

[9] Tim Nugent, Fabio Petroni, Natraj Raman, Lucas Carstens and Jochen L. Leidner, "A Comparison of Classification Models for Natural Disaster and Critical Event Detection from News.", IEEE Big Data DSEM Workshop.

[10] Kunal Chakma, Amitava Das, "A Corpus for Evaluation of Code Mixed Information Retrieval of Hindi-English Tweets", 2016.

[11] M. Anand Kumar, Shivkaran Singh, B. Kavirajan, and K. P. Soman, "Detecting Paraphrases in Indian Languages (DPIL)", FIRE 2016.

[12] Koichi Sato,Junbo Wang,Zixue Cheng, "Detecting Real-time Events using Tweets", IEEE Symposium Series on Computational Intelligence (SSCI) 2016.

[13] Nikhil Dhavase,Prof. A. M. Bagade, “Location Identification for Crime & Disaster Events by Geoparsing Twitter”, International Conference for Convergence for Technology-2014.

Survey on Personality Analysis using Social Media

Sagar Patel, Mansi Nimje, Akshay Shetty and Prof. Sagar Kulkarni

Department of Information technology, Pillai College of Engineering, Navi Mumbai, India - 410206

***Abstract*—** Social media has become a platform for users to present themselves to the world openly by revealing their personal views and insights on their lives. Hence, extracting information from social media and yielding insightful results about the person has become easier. We are beginning to understand that this information can be efficiently utilized to analyze the personality of the concerned person. In this paper, we aim to gain knowledge of the personality of a user by using the social media platform of the concerned user. These social media platforms could be Facebook or Twitter. Personality analysis can help to reveal many types of interactions: it can be used to predict a suitable job for a person and also know about his efficiency in the same; professional, romantic, his nature's traits can also be studied. Personality analysis may even be able to detect the roots of any kind of suspicious, immoral or wrongful trait in a person.

Keywords—Personality, social media, MBTI, Big Five Personality, analysis

1. INTRODUCTION

Personality is defined as the characteristic set of behaviors, cognitions and emotional patterns that evolve from biological and environmental factors. While there is no generally agreed-upon definition of personality, most theories focus on motivation and psychological interactions with one's environment. Trait-based personality theories, such as those defined by Raymond Cattell define personality as the traits that predict a person's behavior. On the other hand, more behavioral-based approaches define personality through learning and habits. Nevertheless, most theories view personality as relatively stable.

Since the inception of social media, a prodigious amount of status updates, tweets and comments have been posted online. The language people use to express themselves can provide clues about the kind of people they are, online and off the digital media. Some personality psychologists study publicly available social media data in addition to solicited surveys. However, they still start with predefined traits like extroversion, neuroticism or narcissism and correlate them with the writing. In other research, linguists have used algorithms to identify topics of conversation, but they do not have much to say about the personalities of the conversationalists. Hence research in this sector is important.

2. LITERATURE SURVEY

1. TwitPersonality: Computing Personality Traits From Tweets Using Word Embeddings and Supervised Learning, 2018 [1] :

Giulio Carducci, Giuseppe Rizzo, Diego Monti, Enrico Palumbo, Maurizio Morisio have proposed a supervised learning approach to compute personality traits by only relying on what an individual tweets about publicly. The approach segments tweets in tokens, then it learns word vector representations as embeddings that are then used to feed a supervised learner classifier. They demonstrate the effectiveness of the approach by measuring the mean squared error of the learned model using an international benchmark of Facebook status updates.

2. Personality Detection by Analysis of Twitter Profiles, 2018 [2] :

Mehul Smriti Raje(B) and Aakarsh Singh, the authors explore the usefulness of Twitter profiles in predicting the personality types of the users. The results of the analysis

of 450 Twitter profiles and over 1 million tweets consist of reviews, comments, personal blogs, feedback, etc.

C. Persona Traits Identification based on Myers-Briggs Type Indicator(MBTI) - A Text Classification, 2018 [3] :

Srilakshmi Bharadwa j, Srinidhi Sridhar, Rahul Choudhary, Ramamoorthy Srinath, their work presents the analysis of text written by a person such as an essay, tweet or blog post and creates a personality profile of the person. The main considerations of the work are the type of data gathered, text preprocessing methods, and machine learning techniques used to estimate personality scores. Various machine learning models and feature vector combinations have been compared, used for deployment of solutions.

D. Personality Prediction of Social Network Users 2018 [4]: *Chaowei Li, Jiale Wan, Bo Wang,* the authors extract social data and questionnaire, and focus on how to use the user text information to predict their personality characteristics. We use the correlation analysis and principal component analysis to select the user information, and then use the multiple regression model, the gray prediction model and the multitasking model to predict and analyze the results.

E. Personality Analysis of the USA Public Using Twitter Profile Pictures, 2017 [5]

Shafaan Khaliq Bhatti, Asia Muneer, M Ikram Lali, Muqaddas Gull, analyze Twitter profile images to predict the personality of major categories of the USA public. For the analysis, we have distributed the USA people into five major categories (Political personalities, Sports Stars, Business Bodies, Hollywood figures, and General public).

F. Personality Classification Based on Twitter Text Using Naive Bayes, KNN and SVM, 2015 [6] :

Bayu Yudha Pratama, Riyanarto Sarno, the authors in this paper talk about the experiment which uses text classification to predict personality based on text written by Twitter users. The languages used are English and Indonesian. Classification methods implemented are Naive Bayes, K-Nearest Neighbors and Support Vector Machine.

G. Personality prediction of Twitter users with Logistic Regression Classifier learned using Stochastic Gradient Descent 2015, IOSR Journal of Computer Engineering (IOSR-JCE) [7] :

Kanupriya Sharma, Amanpreet Kaur, proposed a new approach to predict personality with new insights to predict personality on crucial factors such as scalability and countermeasures to improve the research based on previous work by using a Logistic Regression Classifier with parameter regularization using stochastic gradient descent.

H. Age, Gender and Personality Recognition using Tweets in a Multilingual Setting ",2015, Notebook for PAN at CLEF [8] :

Mounica Arroju, Aftab Hassan, Golnoosh Farnadi, work, describe the properties of our multilingual software submitted for PAN 2015 which recognizes the age, gender and personality traits of Twitter users in four languages, namely, English, Spanish, Dutch and Italian.

I. Predicting Myers-Briggs Type Indicator with Text Classification [9] :

Hernandez, Rayne and Knight, Ian Scott, focus on using machine learning to build a classifier capable of sorting people into their Myers-Briggs Type Indicator personality type based on text samples from their social media posts. Their current system is administered by trained psychologists hover around 0.5.

J. Neural Networks in Predicting Myers Brigg Personality Type from Writing Style [10] :

Anthony Ma and Gus Liu, present a hypothesis that an individual's writing style is largely coupled with their personality traits and present a deep learning model to predict Myers Briggs Personality Type through textual data from books.

3. RESEARCH GAP

1) Most of the authors have done personality analysis on the basis of word-level not at semantic level. [6][7]
2) Future research should look into collecting more sentiment- annotated tweets to get a better handle on the underlying psychological phenomena of opinion and subjectivity.
3) Most of the work done is limited to the English language and hence the involvement of different languages is required. [8]
4) A deeper analysis can be performed to find the intended meaning behind the usage of words.

Weightage can be given to the differences between words depending on their gravity, for example, the words "blue", "sad" and "melancholy" portray different intensities of depression and can make a huge difference during diagnosis. [10]

5) The predictive models must be scalable and dynamic to meet the requirements of ever-growing data and vast possibilities.

4. PROPOSED SYSTEM

4.1 Overview

The proposed system will be designed and trained to accurately identify the personality of the concerned user, using his social media handle. The social media platform chosen for the proposed system is Twitter. Since Twitter is a platform which has data in mostly textual content and it is used by a wide number of people, it is an ideal platform for the system. The end result of the built system will give the personality type of the user according to MBTI personality model. The MBTI model categorizes the personality traits into sixteen types, thus providing a deeper understanding of the personality of the user [13].

4.2 Existing System Architecture

1) Data Collection : According to the survey most of the datasets are built using Twitter API, MyPersonality dataset or datasets from PAN or FIRE [11].
2) Data Preprocessing : In existing systems the data pre-processing includes only tokenization, stopword removal and links removal from their gathered datasets.
3) Feature Extraction : In the existing system the feature extraction work is done on word-level using CountVectorizer, TF-IDF, LIWC and Emolex [3][8]. Also the word count, post length and POST Tagging were used to get features from twitter posts.
4) Training: Currently Naive Bayes, Logistic Regression, Support Vector Machine (SVM) and Xgboost algorithms [6] are in use for their personality classification while few of the authors have implemented the CNN classifier [10] and Stochastic gradient descent algorithm [7] for this task. Also training is performed on a combination of personality traits.
5) Testing : Test was done by considering the accuracy as the model measures metrics on both five big personalities [12] and MBTI personality models [13].

4.3 Proposed System Architecture

The research and survey done have given us an insight on the amount of work currently done in the field of Cognitive Science. Although the survey has enabled us to study the drawbacks and flaws and encouraged us to propose a system of our own which will be more effective and accurate than the present ones. Our proposed system is a complete state by state process including six stages for Personality Analysis. The figure below gives a briefer analysis of the same.

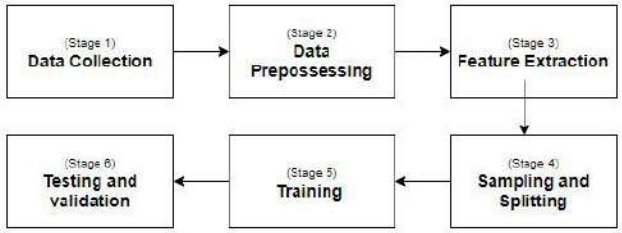


Figure 1 Proposed System

A detailed explanation of the proposed system architecture:

1) Data Collection: In our proposed system we choose to use Kaggle (MBTI) Myers-Briggs Personality Type Dataset [14], since the various other datasets fall insufficient.
2) Data Preprocessing : In this stage we propose to process the data by removing links, HTML Tags, multiple spaces, and stop words. Along with this we will be handling emoticons, morphology and stemming.
3) Feature Extraction: The proposed system will be extracting the features using bag-of-word (BOW), Term frequency-inverse document frequency (TF-IDF), Word2Vec (W2V) and Glove embedding techniques [1].
4) Sampling and Splitting : To balance the disturbed dataset, the concept of upsampling (SMOTE) is used. The resulting dataset will then be split into training dataset (80%) and testing dataset (20%).
5) Training: The dataset will be trained using K-Nearest Neighbour (KNN), Multinomial model,

Logistic Regression [6], Recurrent Neural Network (RNN) with Glow embedding [10] and Support Vector Machine (SVM) [7] for classifying personality according to the MBTI Model [13].

6) Testing and validation: Most of the authors in the existing systems have only considered the accuracy of the model to study its performance. But we propose to consider parameters like accuracy, precision, recall, F-measure and confusion matrix [15] to evaluate the performance of our system.

ACKNOWLEDGMENT

It is our privilege to express our sincerest regards to our supervisor Prof. Sagar Kulkarni 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 to our Head of the Department Dr. Satishkumar Varma and our Principal Dr. Sandeep M. Joshi for encouraging and allowing us to present this work.

REFERENCES

1. Giulio Carducci, Giuseppe Rizzo, Diego Monti, Enrico Palumbo, Maurizio Morisco," TwitPersonality: Computing Personality Traits From Tweets Using Word Embeddings and Supervised Learning, 2018"
2. Mehul Smriti Raje(B) and Aakarsh Singh," Personality Detection by Analysis of Twitter Profiles, 2018"
3. Srilakshmi Bharadwa j, Srinidhi Sridhar, Rahul Choudhary, Ramamoorthy Srinath," Persona Traits Identification based on Myers-Briggs Type Indicator(MBTI) - A Text Classification, 2018"
4. Chaowei Li, Jiale Wan, Bo Wang," Personality Prediction of Social Network Users 2018"
5. Shafaan Khaliq Bhatti, Asia Muneer, M Ikram Lali, Muqaddas Gull," Personality Analysis of the USA Public Using Twitter Profile Pictures"
6. Bayu Yudha Pratama, Riyanarto Sarno," Personality Classification Based on Twitter Text Using Naive Bayes, KNN and SVM"
7. Kanupriya Sharma, Amanpreet Kaur, "Personality prediction of Twitter users with Logistic Regression Classifier learned using Stochastic Gradient Descent 2015, IOSR Journal of Computer Engineering (IOSR-JCE)"
8. Mounica Arroju, Aftab Hassan, Golnoosh Farnadi," Predicting Myers-Briggs Type Indicator with Text Classification"
9. Hernandez, Rayne and Knight, Ian Scott, " Predicting Myers-Briggs Type Indicator with Text Classification"
10. Anthony Ma and Gus Liu, "Neural Networks in Predicting Myers Brigg Personality Type from Writing Style "
11. FIRE, Forum for Information Retrieval Evaluation.
12. https://en.wikipedia.org/wiki/Big_Five_personality_traits, Five Big Personality traits.
13. https://www.myersbriggs.org/my-mbti-personality-type/mbti-basics/home.htm?bhcp=1, Myers-Briggs Type Indicator.
14. https://www.kaggle.com/datasnaek/mbti-type, Kaggle MBTI Dataset.
15. https://scikit-learn.org/stable/auto_examples/model_selection/plot_confusion_matrix.html,Confusion matrix.

E-health chain and anticipation of future disease

Rohit Dhonde[1] , Pradnesh Khedekar[2] , Pradeep Kshirsagar[3] , and Prof. Manasi Kulkarni[4]
1Member, Pillai College of Engineering, New Panvel, Maharashtra – 410206, India
2Member, Pillai College of Engineering, New Panvel, Maharashtra – 410206, India
3Member, Pillai College of Engineering, New Panvel, Maharashtra – 410206, India
4Guide, Pillai College of Engineering, New Panvel, Maharashtra – 410206, India

Abstract—E-Health Chain & Anticipating Future Diseases is a system which aims at maintaining Electronic Health Records (EHRs) in a more efficient way as compared to traditional way of storing and maintaining paper based health records. Digital prescription module can be used by patients to buy medicines from pharmaceutical stores by just providing a unique ID of the patient which helps pharmacists to access the latest prescribed medicines. Electronic Health Records (EHRs) allows doctors to access a patient's health records easily from a single electronic file, doctors can read test results as they are entered, including image files such as X-rays even from remote hospitals. In an emergency situation, a doctor can use a patient's ID code to read time-critical information, such as blood type, allergies, recent treatments. In situations of emergency, the historical data will assist the doctors to take effective actions and safeguard the life of the patients. Tab reminder alert helps patients to take medicines on time as prescribed by doctor. System uses machine learning algorithms to predict and examine future diseases which helps patients to take preventive measures.

Keywords: crime, broken windows, decision trees, classification (linear SVM, Gaussian Naive Bayes), regression (Ridge, XGBoost, KNN , Lasso , SVM, Random Forest, Decision Tree), spatial analysis

I. Introduction

E-Health Chain & Anticipating Future Diseases is a system which aims at maintaining Electronic Health Records (EHRs) in a more efficient way as compared to traditional way of storing and maintaining paper based health records. Digital prescription modules can be used by patients to buy medicines from pharmaceutical stores by just providing a unique ID of the patient which helps pharmacists to access the latest prescribed medicines. Electronic Health Records (EHRs) allows doctors to access a patient's health records easily from a single electronic file, doctors can read test results as they are entered, including image files such as X-rays even from remote hospitals. E-Ambulance is a quick-response solution that can detect and position the phone call for the ambulance and send the emergency ambulance to the necessary point fast. In an emergency situation, a doctor can use a patient's ID code to read time-critical information, such as blood type, allergies, recent treatments. In situations of emergency, the historical data will assist the doctors to take effective actions and safeguard the life of the patients. Tab reminder alert helps patients to take medicines on time as prescribed by doctor. System uses machine learning algorithms to predict and examine future diseases which helps patients to take preventive measures.

II. Proposed Model

A. Overview

The Section presents an Overview and Description of techniques used for the system. E-Health Chain & Anticipating Future Diseases is a system which aims at maintaining Electronic Health Records (EHRs) in a more efficient way as compared to traditional way of storing and maintaining paper based health records.

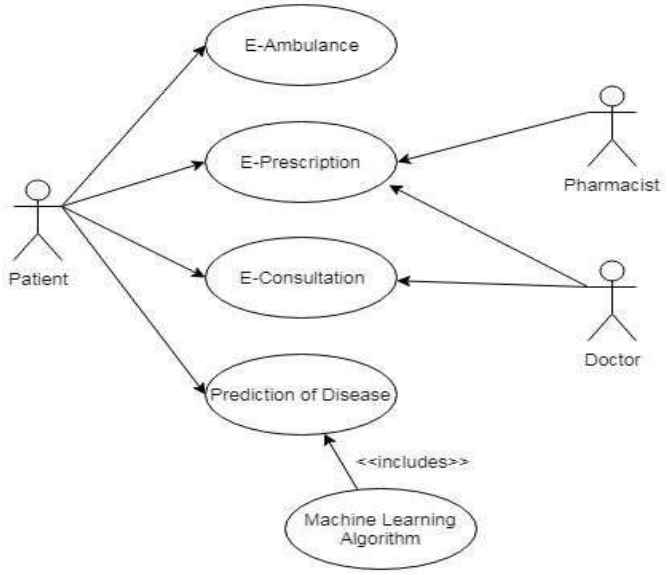


Figure 1: Use Case Diagram

Patient Registration: If Patient is a new user he will enter his personal details and he will have user Id and password through which he can login to the system.

Patient Login: If Patient already has an account then he/she can log into the system.

View Details: Patient and Doctor both can view their information entered in the system. Doctors can also view Patients details and patients can view only doctors with little information.

Diseases Prediction Module: It will ask patients to specify the symptoms caused due to his illness. Module will ask

certain questions regarding his illness and it will do the prediction. The disease based on the symptoms specified by the patient and system will also suggest doctor based on the disease.

Doctor : Doctor will record new data.

Pharmacist : Will provide medicine to the Patient which is prescribed by the Doctor.

Role Based Access Control : Role-based access control (RBAC) is a method of restricting network access based on the roles of individual users within an enterprise. RBAC lets Doctor's have access rights only to the information they need to do their jobs(eg. Patient and Disease Profile) and prevents them from accessing information that doesn't pertain to them.

B. Existing System

The existing manual method of maintaining a patient record, maintains doctor planning data, day to day activities and asking is hard and therefore a system or application which may complete these tasks in a very simple to use is what we are able to attain by this application.

A health system consists of all organizations, folks and actions whose primary intent is to market, restore or maintain health. This includes efforts to influence determinants of health yet as additional direct health-improving activities. A health system is so quite the pyramid of publically closely-held facilities that deliver personal health services. It includes, for instance, a mother caring for a sick kid at home; non-public providers; behaviour modification programmes; vector-control campaigns; insurance organizations; activity health and safety legislation. It includes inter-sectoral action by health employees, for instance, encouraging the ministry of education to market feminine education, a documented determinant of higher health.

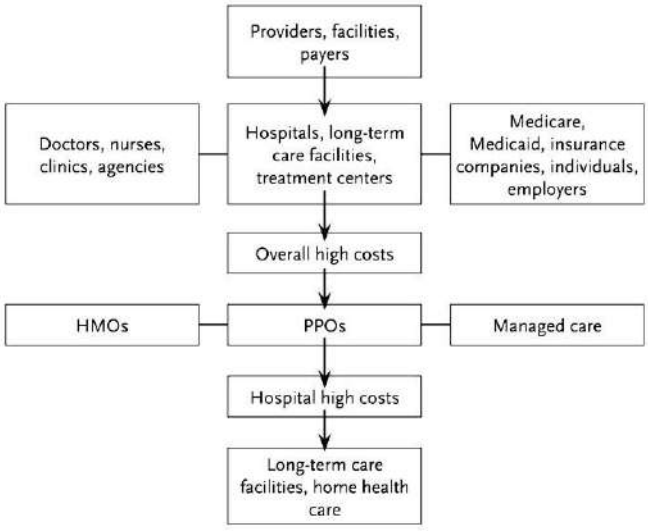


Figure 2: Existing System Architecture

C. Proposed System

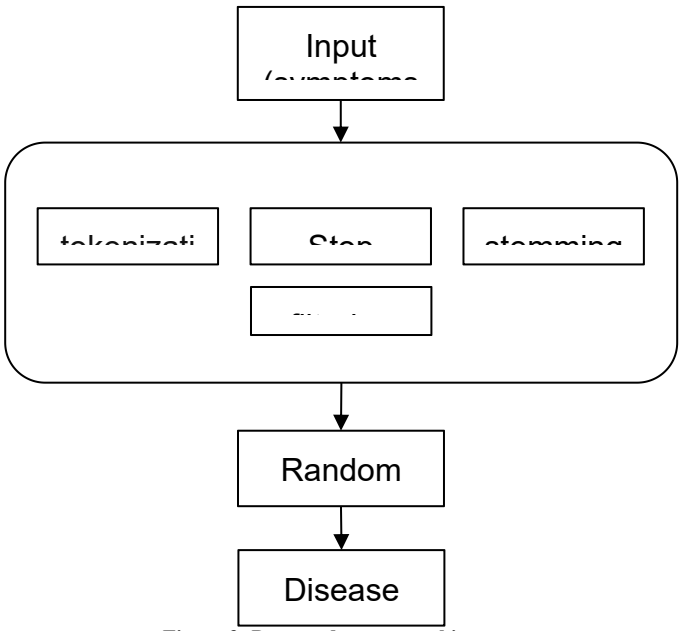


Figure 3: Proposed system architecture

Preprocessing of data :

1. **Punctuation Removal** : The punctuation marks are removed from the text because they add no meaning to the data thus of no use.
2. **Blank/White space Removal** :To remove leading and ending spaces, you can use the *strip()* function. It is helping to reduce the memory uses and increase the efficiency of the model.
3. **Stemming/Lemmatization:** The aim of lemmatization, like stemming, is to reduce inflectional forms to a common base form. As opposition stemming, lemmatization doesn't merely lop off inflections. Instead it uses lexical information bases to urge the right base types of words.
4. **Stop-Word Removal:** The removal of Stopwords (such as "is","the".etc) is called Stop Word Removal. Stopwords add very little meaning so if removed the database space is saved and processing speed improves.

Disease Prediction Algorithm

Random Forest:

Random forests or random decision forests or associate ensemble learning methodology for classification, regression and different tasks that operates by constructing a large number of decision trees at training time and outputting the class that's the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees habit of overfitting to their training set.

In decision prediction the Random Forest algorithm is used to process the symptoms as input and gives diseases as output. Based on the probability of predicted diseases top 4 diseases with brief information are shown as output.

D. Hardware and Software Specifications

The experiment setup is carried out on a computer system which has the different hardware and software specifications as given in Table 1 and Table 2 respectively.

Table 1: Hardware details

Hardware	Details
RAM	512MB or above
Hard disk drive	500 MB or above (1GB or more recommended)
Processor	Pentium IV or above
Input Device	Standard Keyboard and Mouse
Other	Computer, Laptop

Table 2: Software details

Operating System	Windows 7/8/10, Linux etc.
Programming Language	HTML5, CSS3, Bootstrap, Javascript, Ajax, Php, Python.
Framework	Laravel and Flask.
Database	MySql
Browser	Chrome

III. PROJECT INPUTS AND OUTPUTS

A. Input Details

Datasets has been taken from the healthcare organization like MayoClinic, National Health Service, National Health Portal of India and many other healthcare websites.Datasets contains information about diseases and symptoms.Datasets also contains information about the disease details.We have taken data of Medical Conditions and Symptoms which patients have suffered.

Table 3: Dataset statistics.

Dataset	Link	Format
National Health Service (NHS)	https://www.nhs.uk/conditions/	[disease, symptom]
National Health Portal (NHP)	https://www.nhp.gov.in/disease-a-z	[disease, symptom]
Mayo clinic	https://www.mayoclinic.org/symptoms	[disease, symptom]
Presbyterian Hospital data	http://people.dbmi.columbia.edu/~friedma/Projects/DiseaseSymptomKB/index.html	[disease, count of disease, symptom]

In table 3, the Format column shows how the data is stored in the CSV. Each data is scraped from a legitimate website given in link column and stored after cleaning of data in CSV format.

B. Data Processing and Evaluation

Preprocessing of data :
Library used for preprocessing of data : NLTK
NLTK:

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces. Natural Language Processing with Python provides a practical introduction to programming for language processing. Written by the creators of NLTK, it guides the reader through the basics of writing Python programs, operating with corpora, categorizing text, analyzing linguistic structure, and more.

1. **Punctuation Removal** : The punctuation marks are removed from the text because they add no meaning to the data thus of no use.

The following code removes this set of symbols [!"#$%&'()*+,-./:;<=>?@[\]^_`{|}~]

example:
INPUT ="Detecting Key- Needs in Crisis."
OUTPUT="Detecting Key Needs in Crisis"

2. **Blank/White space Removal** :To remove leading and ending spaces, you can use the *strip()* function. It is helping to reduce the memory uses and increase the efficiency of the model.

example:
INPUT =" \t Blurred and distorted vision \t"
OUTPUT="blurred and distorted vision."

3. **Stemming/Lemmatization:** The aim of lemmatization, like stemming, is to reduce inflectional forms to a common base form. As opposition stemming, lemmatization doesn't merely lop off inflections. Instead it uses lexical knowledge bases to induce the proper base forms of words.

example:
INPUT="Sleep disturbances and headaches"
OUTPUT="Sleep disturbance and headache"

4. **Stop-Word Removal:** The removal of Stopwords (such as "is","the".etc) is called Stop Word Removal. Stopwords add very little meaning so if removed the database space is saved and processing speed improves.

example:
INPUT=["Incompetent","in", "lips"," closure."]
stopwords=['in','.']
OUTPUT=["Incompetent", "lips"," closure."]

By using the NLTK code all the disease symptoms are cleaned for training as it removes HTML tags, stopwords and special characters.

Evaluation Parameters Details

Algorithm used: Random forest

Example:

```
from sklearn.ensemble import RandomForestClassifier
clf_rf=RandomForestClassifier(n_estimators=100)
```

Split data for training and testing purposes:

```
Example : x = df_pivoted[cols]#symptoms
          y = df_pivoted['disease']
        x_train, x_test, y_train, y_test = train_test_split(x, y,
        test_size=0.33, random_state=42)
```

Prediction using multinomial naive bayes model:

```
Example : disease_pred = clf_rf.predict(x_test1)
```

The accuracy of a Disease predictor can be evaluated by using : sklearn.metrics

```
Example : from sklearn.metrics import accuracy_score
          accuracy_score(y_test1, y_pred)
Output :  0.9841269841698
```

C. Output Details

Web application :

We have deployed our project using the Flask framework of Python and hosted a virtual environment provided by heroku. http://node-179.herokuapp.com/ In web application we are taking a list of symptoms from the user. In the backend the passed symptoms are evaluated as id stored in the dictionary as input in symptom_list. array of symptoms is passed as input for prediction in multinomial naive bayes model. It has only 0&1 value for each symptom. model.predict() gives disease which has the highest probability. We find the top 4 diseases which have better probability than other diseases present in targets using model.predict_proba() function. In below Figure.4 we are getting diseases : [‘herpes simplex eye infections’, ‘whooping cough/Pertussis’, ‘Benign Essential Biepharospasm’, ‘Astigatism’] .

Screenshots

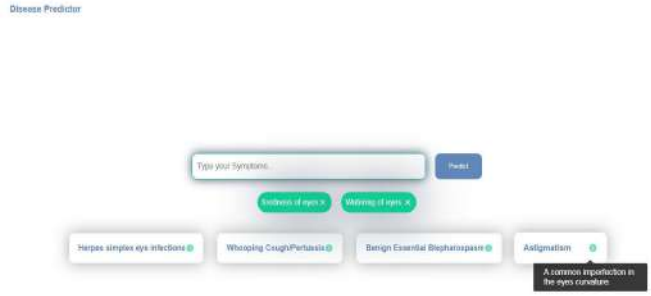


Figure 4: Disease Predictor

Diseases prediction module will ask patients to specify the symptoms caused due to his illness. Module will ask certain questions regarding his illness and it will do the prediction. The disease based on the symptoms specified by the patient and system will also suggest doctor based on the disease.

Homepage :

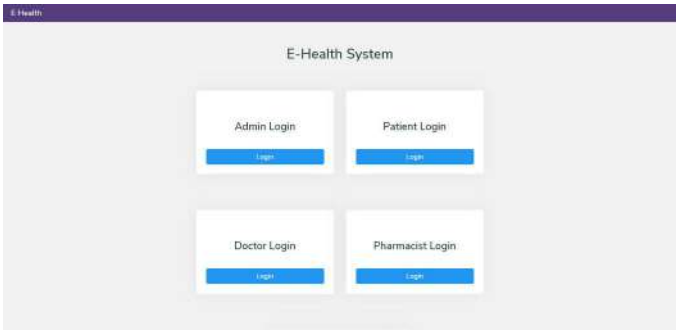


Figure 5: homepage

This is the first screen of our system through which various users and system administrators can log in into the E-Health chain system.

Figure 6: login page

The login screen for users of the system.

Admin module

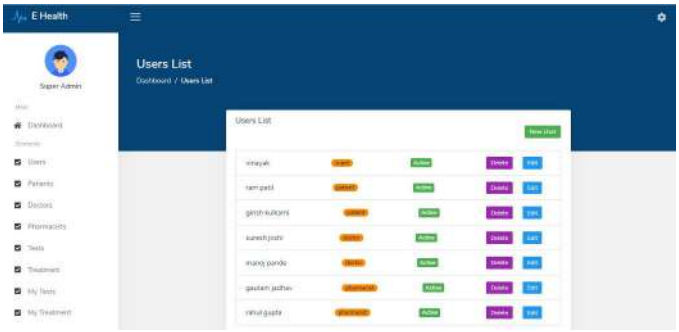


Figure 7: admin module

This the Dashboard for the System administrator of the E-health chain where there are various options also it has authority to add users in various roles for systems like Patients, Doctors and Pharmacists. Also can view various data related to various users of the system.

Patient module

This is the Dashboard for Patient where the E-health records are stored which he can view and consult a doctor in any emergency at any time any place also it helps to keep track of various ongoing and previous treatments.

Figure 8: Patient module

Pharmacist module

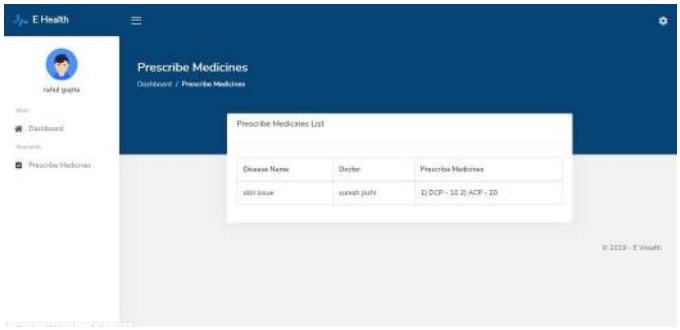


Figure 9: Pharmacist module

This is the Dashboard for the Pharmacist through which he or she can view only prescribed medicines by a specific doctor for specific patients through a unique identification number.

Doctor module

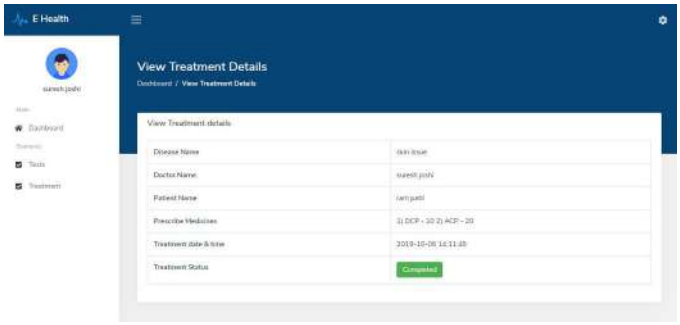


Figure 10: Doctor module

This is the Dashboard for the Doctor through which doctors can add various information related to patients like treatments and prescribed medicines also it maintains the status of treatment.

IV. Conclusion

In this report, the study of different domain techniques is presented. Traditional styles of software systems did not meet the necessities of our system. We thus based our design upon a service-oriented architecture that can satisfy the stated functional requirements. Our Ehealth portals can integrate different medical services and applications. In our system there are various modules like E-health records, E-Prescription through which various services have been implemented and integrated into our Ehealth portal. We also pointed out limitations found in the access control module of many off-the-shelf software components. Our answer was supported by a two-tier access management design that integrated existing RBAC modules with a role-based access management extension. This style geneticized the benefits of each model and was cost-effective. We conjointly indicated and projected solutions for the inconsistency issue at intervals the two-tier model, the predicate attribute naming inconsistent issue, and also the restricted attribute issue in applying our model.

The "anticipation of future disease" module is implemented using Machine learning Random Forest algorithm by using structured and unstructured data from legitimate websites. Compared to several typical calculating algorithms, the scheming accuracy of our proposed algorithm reaches 98.41%. The results consist of the possibility of occurrences of diseases.

V. Future Scope

Implement E-Ambulance module, which is a quick-response solution that can detect and position the phone call for the ambulance and send the emergency ambulance to the necessary point fast. The historical data will assist the doctors to take effective actions and safeguard the life of the patients.

Making this system as one of the government portals like nhp. Creating an easy to use mobile application for remote use. Increase the accuracy and correctness of the disease prediction module. Also increasing the count of diseases can be predicted using the disease prediction module.

References

[1] Machine learning applications in cancer prognosis and prediction by Konstantina Kourou,Themis P. Exarchos, Konstantinos P. Exarchos Michalis V. Karamouzis, Dimitrios I. Fotiadis

[2] Predictive Analytics for Chronic Kidney Disease Using Machine Learning Techniques by Anusorn Charleonnan, Thipwan Fufaung, Tippawan Niyomwong, Wandee Chokchueypattanakit, Sathit Suwannawach, Nitat Ninchawee

[3] Performance Analysis of Machine Learning Algorithms on Diabetes Dataset using Big Data Analytics by P. Suresh Kumar, S. Pranavi

[4] Disease Prediction by Machine Learning Over Big Data From Healthcare Communities by MIN CHEN , (Senior Member, IEEE), YIXUE HAO1 , KAI HWANG , (Life Fellow, IEEE), LU WANG, AND LIN WANG

[5] Applying Machine Learning Techniques for Predicting the Risk of Chronic Kidney Disease by K. R. Anantha Padmanaban* and G. Parthiban

[6] PREDICTION OF PROBABILITY OF DISEASE BASED ON SYMPTOMS USING MACHINE LEARNING ALGORITHM by Harini D , Natesh

[7] Data Analysis on Health Management Systems for Improving Doctor's Advice on Patients by Li-jia Yu, Hua-qiong Wang , Ling Gou1 , Yu Tian , Jing-song Li

[8] Heart Disease Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm

Optimization and Ant Colony Optimization by Youness Khourdifi, Mohamed Bahaj.
[9] DISEASE PREDICTION BY USING MACHINE LEARNING by Sayali Ambekar, Dr.Rashmi Phalnikar.
[10] LIVER DISEASE PREDICTION BY USING DIFFERENT DECISION TREE TECHNIQUES by Nazmun Nahar and Ferdous Ara.

INTELLIGENT TRAFFIC INFORMATION SYSTEM BASED ON INTERNET OF THINGS

[#1]SOURABH KULKARNI
STUDENT(IT)
PCE, NEW PANVEL
kulkarnisrit16e@student.mes.ac.in

[#2]SHREYA NAYAK
STUDENT(IT)
PCE, NEW PANVEL
nayaksjit16e@student.mes.ac.in

[#3]SAGARIKA CHANDEL
STUDENT(IT)
PCE, NEW PANVEL
chandelsasaet16e@student.mes.ac.in

SHEETAL GAWANDE
FACULTY
PCE, NEW PANVEL
sheetalp@mes.ac.in

Abstract:

In recent years, the popularity of personal vehicles is a major issue in big cities which causes traffic congestion, environmental pollution, waste of time and much more. Mostly, traffic congestion causes accidents. Hence, traffic management is the vital issue in big cities. Manual traffic control by policemen as well as the predefined set time for the signal at all circumstances has not proved to be efficient. The model makes use of cloud for delivering different services such as server To address the above-mentioned issues, this paper proposes an internet-of-things (IoT) based model which also gives priority to emergency vehicles and avoiding dumping areas. This model makes use of cloud to deliver services such as storage, application. With the help of cloud computing, we can store the data on the internet which gives continuous update so that it can handle the traffic smoothly. A real time traffic information collection and monitoring system is proposed which solves the problem of real time monitoring and controlling road traffic. This system employs key technologies: Internet of Things, Load Cells and RF Transmitters and Receivers to collect, store, manage and supervise traffic information. Advantages of this model are cost effectiveness, fuel efficiency and reduced travelling time.

Keywords:

IOT, Raspberry Pi 3, RFID, Load Cell, MQ2 Gas Sensor

Introduction

IoT is nothing but the network of interconnected things/devices embedded with sensors, software, network connectivity and required electronics. Such an interconnection makes it possible for sensors to be responsive by collecting and exchanging data. A sensor, which is useless by itself, plays a very important role when used in an electronic system. A sensor measures physical phenomenon in the environment such as temperature and transform it into an electronic signal. Various types of sensors are required for a variety of applications.

Modern transport system fails to provide smooth transportation to citizens in the world of continuous and fast paced development. This eventually leads to excessive traffic jams which result into delays professional and personal spots. In addition to this various other problems also arise such as mental frustration causing road rages, fuel wastage, and wear and tear of vehicles. Nowadays, traffic issues are eventually faced by everyone due to an increase in the number of vehicles. IoT can be used to resolve the problem of traffic congestion many types of sensors can also be used. When there are less number of vehicles then the traffic signals should be turned off so that the vehicles can go freely whereas in crowded areas the traffic signals should work properly to avoid the chaos of traffic and jam. Also other sensors can be used to detect the dumping areas so that the user can avoid that area also the smoky areas can be avoided by using sensors. The vehicle which is heavy in weight and suppose that bridge doesn't have the capacity to take it

then the sensor will give an alert message to the user so that they can avoid that bridge and search an alternative necessary path..

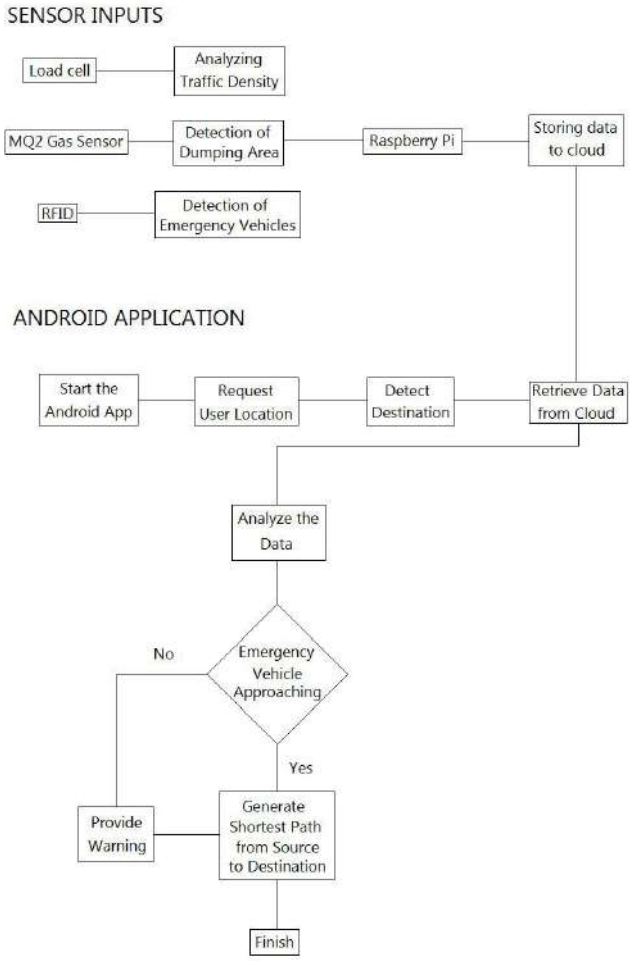


Fig 1.1 Working of Traffic Information System

Literature Survey

Pallavi Belokar and Prof. Kavita Joshi, in their paper "Intelligent Traffic Management Control Systems" design a model integrated with GPS-GSM in which the traffic information is imparted to the drivers via SMS which can help them to choose traffic avoiding route to the destination. Here, the driver receives real time traffic information from the server after manually enquiring for it through GSM. The real-time traffic information is collected through IR sensors. The proposed model software is deployed in the keil micro vision compiler environment using programming language "Embedded C".[1]

"IoT based Intelligent Traffic Control System" by Harshini Vijetha H and Nataraj K R, proposed a new approach of Controlling Traffic System which makes use of Pi-Camera and Ir sensor for detecting traffic density. This model uses RFID for confirming zero traffic region for emergency vehicles and also for tracking stolen mobile phones. It uses a dual mode control system, automatic mode and manual mode. The automatic mode does not involve human intervention whereas manual doe involve human intervention to manually search and find the route.[2]

"Intelligent Traffic Information System Based on Integration of Internet of Things and Agent Technology" by Hasan Omar Al-Sakran put forward an architecture that integrates IoT with agent technology into a single platform where the agent technology handles effective communication and interfaces among a large number of heterogeneous highly distributed and decentralized devices within the IoT. It presents a framework distributed traffic simulation model within NetLogo, an agent-based environment for IoT traffic monitoring system using mobile agent technology.[3]

I. Made Oka Widyantara and Nyoman Putra Sastra, in their research paper, "Internet of Things for Intelligent Traffic Monitoring System: A Case Study in Denpasar", intended to determine the design of the implementation of the IoT for Intelligent Traffic Monitoring System (ITMS) in Denpasar city, Bali, Indonesia. The main goal of this research was to visualize the traffic on Web-based GPS/GPRS. Their implementation mainly focuses on acquisition of traffic by leveraging the capabilities of GPS as a sensor, GPRS based data transport, and the design of a Web/GIS based traffic monitoring software.[4]

This paper highlights the optimization of traffic data collection in a city using sensors and microcontroller. The paper provides configuration to minimize the possibilities of traffic jams problems. It is observed due to this proposed system of Intelligent Traffic data collection is more efficient and convenient, more

distance covered by average vehicles and efficient operation during emergency mode.[1] Identification of heavy traffic areas is done successfully and fast. Clearance of traffic for an emergency automobile is successfully implemented. Hence, many precious lives would be saved.[2] The proposed system can provide a new way of monitoring traffic flow that helps to improve traffic conditions and resource utilization. In addition, transport administration department, using real-time traffic monitoring information, can in time detect potentially dangerous situations and take necessary actions to prevent traffic congestion and minimize the number of accidents thus ensuring the safety of road traffic.[3]

Proposed Methodology

Now a days there is an increase in vehicle and due to our busy life, we prefer to travel by our private vehicle rather than traveling by public transport and due to this there is an increase in traffic problems. In our proposed system we have provided solutions for traffic congestion and monitoring system. In this we are providing an android application for the user where he can detect his live location first then he can search for the shortest path as well as alternative path where he can avoid traffic jams. We are also using sensors for detecting the dumping areas and smoky area of danger will also be included in the application so that the user can avoid this zone.

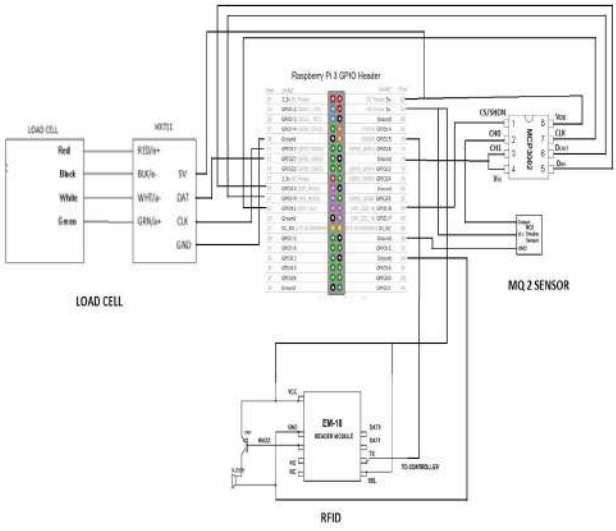


Fig 1.2 Circuit Diagram

Implementation

The load cell, which will fit in the road, will analyze the traffic passing over it. Predefined threshold values and range of weight categories will differentiate the type of traffic in the region. The data collected will be uploaded to the cloud storage.

Further, MQ2 sensor placed at the dumping areas will detect the methane gas emitted from the garbage. More amount of methane gas will indicate that the garbage containers are completely filled and there is a possibility that the garbage is spilled out on the road. Hence this data is collected and uploaded to the cloud database.

RF identification technology is used for creating alert about emergency vehicles. RF transmitters are placed in the emergency vehicles such as Ambulance, Fire Brigades and Police Vans. RF readers will be placed in dividers or on the signal pole. As soon as any emergency vehicle enters within the region of RF readers then an alert will be created which will be displayed on the user application. This data is also sent to cloud database.

All this data is analyzed and fetched in an android application. In this Android application, the user will be able to find his/her current location and find the shortest path to the destination. For suggesting shortest path we will be using Dijkstra's algorithm. Dijkstra's algorithm (or Dijkstra's Shortest Path First algorithm, SPF algorithm) is an algorithm for finding the shortest paths between nodes in a graph, which may represent road networks. While suggesting shortest path, heavy traffic areas and garbage dumping areas are considered and mostly avoided for smooth and hassle free journey from the user location to the destination.

Application

(i)Dynamic traffic light sequence:- RFID for dynamic traffic light sequences circumvents or avoids problems that usually arise with systems that use image processing and beam interruption techniques. RFID technology with appropriate algorithm and database can be applied to a multi-vehicle, multi-lane and multi-road

junction area to provide an efficient time management scheme. A dynamic time schedule can be worked out for the passage of each column. The simulation showing the dynamic sequence algorithm can adjust itself even with the presence of some extreme cases. This system will be able to emulate the judgment of a traffic police officer on duty, by considering the number of vehicles in each column and the routing properties.

(ii)Finding shortest path and other areas of traffic:- Android application provides the shortest and alternative path along with the dumping areas and the smoke areas where there is danger so that the user can avoid that place.

Conclusion

Fig1.3: MQ2 Sensor Cloud Data

The continuous growth of population all over the world create great challenges to the transport management system. Due to an increase in the number of vehicles, it is necessary to take effective steps in order to control the traffic and hence avoid all types of loses that is caused due to traffic. The study aimed at understanding the traffic issue and recommending improvements to facilities smoother traffic flows. The development of the control system to deal with traffic congestion in urban areas is a critical issue. Not only this but our project highlights the problem as well as the solution, for example, we specify whether there is traffic congestion or not, but we also provide alternate paths for the same destination .along with alternative paths there is an alert message about the garbage area for the user for a smooth journey.

Reference

1. Ms. Pallavi K. Belokar, Prof. Kavita V. Joshi, “Intelligent Traffic Management Control Systems” International Journal of Innovative Research in Advanced Engineering (IJIRAE), Volume 1 Issue 5 (June 2014)
https://ijirae.com/images/downloads/vol1issue5/JNEC10088(30).pdf

2. Harshini Vijetha H, Dr. Nataraj K R, “ IOT Based Intelligent Traffic Control System” International Journal for Research in Applied Science & Engineering Technology (IJRASET), Volume 5 Issue V, May 2017, IC Value: 45.98
https://www.ijraset.com/fileserve.php?FID=7713

3. Hasan Omar Al-Sakran, “Intelligent Traffic Information System Based on Internet of Things and Agent Technology” (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 6, No. 2, 2015
https://thesai.org/Publications/ViewPaper?Volume=6&Issue=2&Code=ijacsa&SerialNo=6

4. I. Made Oka Widyantara, Nyoman Putra Sastra, “Internet of Things for Intelligent Traffic Information System: A Case Study in Denpasar”, International Journal of Computer Trends and Technology(IJCTT) - volume 30 Number 3 - December 2015
https://www.ijcttjournal.org/archives/ijctt-v30p130

Virtual Assistant for Visually Impaired

Vipul Sharma[1], Vishal Mahendra Singh[2], Sharan Thanneeru[3], Prof. Amol Kharat[4]
[1,2,3,4]Department of Information technology, Pillai College of Engineering, University of Mumbai
[1]vipulsit16e@student.mes.ac.in
[2]vishalmasin16e@student.mes.ac.in
[3]thanneerushveit16e@student.mes.ac.in
[4]akharat@mes.ac.in

***Abstract*- The field of artificial intelligence has led to various virtual assistants such as Siri in iPhone, Google Allo, Microsoft Cortana, and so on. Even after such progression, very little has been done to implement these technologies to assist the visually impaired community. Recognizing a person or distinguishing an object, these tasks are straightforward for common people but can be very difficult for people that are partly or completely blind. Their lives can be made smoother by assisting them to detect what is present in front of them at that instant. We aim to develop a system/assistant that will serve to guide a visually impaired person and will indicate the person by speaking through the earpiece. The system will help the person recognize people, add new faces and detect objects that are in their vicinity. We will have a mobile application which will consist of numerous deep learning models that will help applications increase its administration. The primary working of the system will consist of the camera continuously feeding images for inputs, the core system processing this input information and the earpiece acting as the output device to provide this output to the user.**

***Keywords*- Face Recognition, Object Detection, Cognitive Services, Text-to-Speech, Deep Learning.**

I. INTRODUCTION

"Virtual Assistant for visually impaired", the said project applies the concept of Deep learning i.e. Neural networks. The models employed for our project are - Face Detection and Object Detection. The system comprises a camera that acquires images and sends them to the application, where a powerful processor derives information from them and explains them to the user through a distinct audible message. The device will continuously detect all the faces in front of the person and verify them against all the faces of the people who have been previously taught to the device.

II. PROPOSED SYSTEM

In the system level, we could say that the novelty lies in the real-time web application. The already existing system comprises modules such as Image processing, Speech processing, etc, therefore the problems faced by blind people are often reduced to a particular extent. But neither are these modules enough nor are they implemented purposefully such that they assist the visually impaired. Taking these limitations into consideration, the system we have developed overcomes these drawbacks and helps build a system that assists the needful in a better and more appropriate manner.

Modules focused upon by us:-

A. ***Text-to-Speech***

This module comprises text and speech processing. The main purpose of this module is to take into consideration all the text provided and convert these into the appropriate audio output using speech processing. We have implemented a dynamic system that makes use of Google API (Gttx) for the conversion of Text to Speech dynamically provided that good internet connectivity is present.

B. ***Object Recognition***

Object Recognition is a process in which Real-world objects are identified using Image processing. It is an important operation that will aid visually impaired to locate their frequently used day to day objects. The system that we have developed provides support in visual aid by assisting to dynamically locate and identify the objects in an image and providing the text output for the same.

C. ***Face recognition***

Some face recognition algorithms identify countenance by extracting landmarks, or features, from a picture of the subject's face that includes the features shape of the jaw, nose, cheek, facial hair and other such characteristics. The features of the image in consideration are then compared with other images having similar features. The algorithm normalizes a dataset of face embeddings then compresses these embeddings, only saving the information within the image that's useful for face recognition. Eventually what we will be obtaining is a bounding box surrounding the face in the live monitoring having the name of the person and the confidence attached to the bounding box.

III. IMPLEMENTATION

The system developed is deployed on the web as a website. The website is built on the backbone of flask, which serves the purpose of providing connectivity between the python code and the HTML.

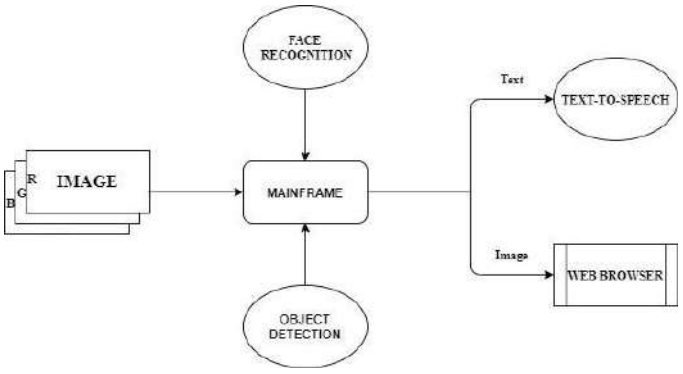


Fig 1. Implementation flow.

When the website is loaded, the object detection module starts its processing and the objects detected by this module are displayed on the page as well as delivered to the user via an earpiece/speaker. Along with this, we also have two buttons ('Switch to Face' and 'Stop') on the landing page that are well separated to be easily accessible. Clicking on 'Stop' results in pausing the Livestream until the 'Start' button is clicked. The 'Switch to Face' button on click will switch to the page where the Face Recognition processing begins. We have also included the buffer which can only contain a maximum of five entities (objects/people) at a time. Each entity will be converted to speech in every 20 seconds if it still exists in the frame.

The 'Face Detection' module is implemented similarly as the 'Object Detection' module using the same layout for the buttons. Here, the clicking of the 'Stop' button will have the same function as mentioned above whereas a click on the 'Add Face' button will capture the current frame and prompt the user to speak out the name of the person whose face is being added. The name is spoken into the microphone by the user and the speech-to-text model converts this audio into the text and stores the text with the captured frame into the database. All the processing is carried out in the python engine and is displayed using HTML to the user. Thus implementing all these, we obtain a system that is more relevant and more assistive to the user.

IV. RESULT

The system is deployed as a web application which, when opened on any mobile browser, gives us the landing page shown below. Along with the landing page, we have two additional pages that play an important role in our system and play a fundamental role in its deployment. When the user presses any of these buttons, the command will be addressed to the user via earpiece/speaker. All these buttons are large in size and are separated properly, so that it is convenient for the visually impaired user to distinguish between them.

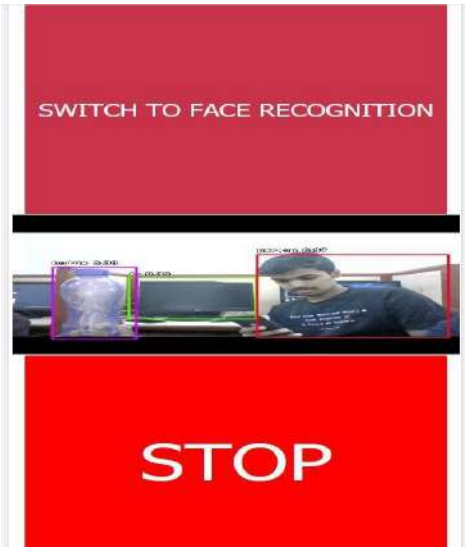


Fig 2. Landing Page - Object Detection.

This page consists of two buttons- one at the top and the other at the bottom of the page. We have a block in the center of the page which provides a continuous live stream that is displayed through the phone. Above the block is the "SWITCH TO FACE RECOGNITION" button which, when clicked, deploys the Face recognition model and directs the user to that page. The other button is named "STOP" and resides below the live stream block. When clicked, this button will stop the current processing model and redirect the user to the page having the "START" button.

Fig 3. Face Recognition Page.

A click on the "SWITCH TO FACE RECOGNITION" button, the system is directed to a new page where the face recognition functionality begins its execution. Similar to the landing page, this page consists of two buttons- one at the top and the other at the bottom of the page. The button at the top is named "ADD FACE" whilst the button at the bottom of the page is named "STOP". If an unknown face is encountered, we can click on the "ADD FACE' button at the top of the page to add the unknown face into the Facial database. The "STOP" button executes the same functionality as before and will stop the current processing model and redirect the user to the page where the "START" button resides. The block in the center of the page separates the two buttons and continues to provide the live stream and displays it through the phone's browser window. All the faces recognized in the live stream are addressed to the user via earpiece/speaker.

Fig 4. Page to START the system after it is stopped.

This page consists of a single large button named "START". When the "STOP" button on either of the Face Recognition page or the Object Detection page is clicked, the user is redirected to this page where the "START" button resides.This enables the user to start the system anew after it has been stopped. Hence, allowing the user to begin the system according to their convenience and usability. This project is available on https://github.com/Deimos-M/DL-Virtual-Assistant.

V. CONCLUSION

In this paper, various techniques to implement the aforementioned system are analyzed and summarized. Different systems have different ways of implementation along with some limitations and restrictions. These types of systems are very critical for multiple reasons and the occurrence of an error in such a system/device may cause catastrophic damage and loss. The system we are achieving overcomes the limitations of the already implemented systems. Our system consists of a basic UI on a web-based application and comprises several Deep learning models; some of them are object detection, face recognition, speech recognition and so on. These modules will work together and assist in vital activities like object detection as well as face detection and recognition for the visually impaired.

VI. FUTURE SCOPE

There are various applications of this domain system. The future scopes are listed below.

A. Alerting the visually impaired person about the Obstacle Position

We would implement the device in such a way that the sensors will be mounted on a spectacle and this would help the person wearing the spectacle detect the obstacle position in front of their vision in the walking path.

B. Voice Command and Emergency Voice Call Establishment

We would include the facility to save an emergency number in the application so that the visually impaired person can establish a voice call to the predefined number by using his/her voice command. When the visually impaired person wants to give a voice command, he/she need not touch the phone and just pressing the lock button thrice on the phone will lead to prompt command and by uttering “HELP” this voice command will connect through a voice call to a predefined number.

C. Text Reader

This system will help a visually impaired person to listen to the text which is written in any literature or any book. The system will take a pic and it will recognize the text written on it using image processing. This recognized text is then converted to speech using a text-to-speech model.

REFERENCES

1. Marcos Barata, Afan Galih Salman, Ikhtiar Faahakhododo, Bayu Kanigoro,"Android based voice assistant for blind people", Library Hi Tech News, Vol. 35 Issue: 6, pp.9-11 (2018).
2. Md. Siddiqur Rahman Tanveer, M.M.A. Hashem and Md. Kowsar Hossain , “ Android Assistant EyeMate for Blind and Blind Tracker” (2018).
3. Joseph Redmon, Ali Farhadi, “An Incremental Improvement”, Pjreddie (2018).
4. Vincent Gaudissart, Silvio Ferreira, Céline Thillou, Bernard Gosselin,“Mobile Reading Assistant for Blind People” (2018).
5. DR. Kavitha C, MR. Nithin V Gopal, MS. Nidhi Amarnath, MR. Prajwal G, MS. Supreetha R.R., “VIRTUAL ASSISTANT FOR BLIND PEOPLE” (2018).
6. Prof. Priya U. Thakare, Kote Shubham, Pawale Ankit, Rajguru Ajinkya, Shelke Om, “Smart Assistance System for the Visually Impaired” International Journal of Scientific and Research Publications, Volume 7, Issue 12, December 2017 378 ISSN 2250-3153 (2017).
7. Faizan Ahmad ,Aaima Najamand Zeeshan Ahmed, “ Image-based Face Detection and Recognition” ,Arxiv (2015).
8. R. Velázquez, Wearable Assistive Devices for the Blind. Chapter 17 in A. Lay-Ekuakille & S.C. Mukhopadhyay (Eds.), Wearable and Autonomous Biomedical Devices and Systems for Smart Environment: Issues and Characterization, LNEE 75, Springer, pp 331-349 (2010).

Mood Based Music Player

Sajida Begum*, Sakshi Manjari*, Pranali Sawant*, Gayatri Hegde**

***Student - Information Technology, Pillai College of Engineering**

****Faculty - Information Technology, Pillai College of Engineering**

Abstract- Facial expressions are one of the best ways to determine how a person is feeling currently. The aim of the project is to generate a website where user's mood will be identified based on their facial expression and music will be generated according to their emotional state. The project consists of two models: 1. Emotion detection model and 2. Music generation model. In emotion detection model , facial expressions will be identified and classified into seven sentiment categories : "Happy, Sad, Angry, Neutral, fear, Sad, Surprised and Disgust" using CNN(Convolutional neural Network). In music generation model, music will be generated according to the identified emotion using LSTM architecture (Long Short- Term Memory). LSTM is a recurrent neural network (RNN) architecture that remembers values over arbitrary intervals.

I. INTRODUCTION

Music plays a very important role in everyone's life as its one of the important sources of entertainment as well as a way to help in dealing with one's present emotional state. Sometimes it can be very time consuming and hectic to find music depending upon the current mood of an individual. Hence, we have developed a mood-based music player where music will be generated according to the emotional state of the user. To determine the emotional state of a person, facial expressions are one of the best ways as it reflects how the person is feeling currently. In this paper, firstly, we capture the facial expressions of the user as it is one of the best ways to determine how the person is feeling currently using Convolutional Neural Network (CNN). The facial expressions are categorized into 7 categories - "Happy, Sad, Angry, Neutral, fear, Sad, Surprised and Disgust" Secondly, music is generated depending upon the captured expressions that indicates certain emotions using Long - Short Term Memory (LSTM).

II. LITERATURE SURVEY

It has been observed that there are majorly two methods for Emotion Detection (Image classification) : using CNN (Convolutional Neural Network) and SVM (Support Vector Machine).In Music generation, method varied depending on the type of music files used as dataset such as WAV, MIDI, ABC, etc. (i) The paper SentiMozart: Music Generation based on Emotions [2] represents the project in two parts: First, emotion of people is captured from their images and is categorized into 7 major categories: Angry, Disgust, Fear, Happy, Sad, Surprise and Neutral using Convolutional Neural Network (CNN). Second, music is generated based on these emotions using Long Short-Term Memory (LSTM). (ii) The paper MoodyPlayer: A Mood Based Music Player [1]represents following stages: in first stage, face detection is done from an image, for this various techniques are used such as model based face tracking which includes real-time face detection using edge orientation matching, Robust face detection using Hausdorff distance , weak classifier cascade which includes Viola and Jones algorithm, and Histograms of OrientedGradients (HOG) descriptors. In the next stage, features are extracted from the detected face. (iii) **A Survey: Expression Based Music Player [5]**, this paper deals with connecting emotion of the user along with music systematically. Expression based Music Player involves the image processing, facial feature detection, expression classification and audio feature extraction. In many research papers on Emotion Detection with Music Generation, the biggest problem with the systems was manual setting of the user's emotion. After the survey it has been observed that for emotion detection, CNN performed much better than SVM. Additionally, SVM required many other steps including image processing, face detection, facial feature extraction, etc.

III. IDENTIFY, RESEARCH AND COLLECT IDEA

We reviewed some implementation techniques like SVM (Support Vector Machine), CNN (Convolutional Neural Network) for emotion detection and RNN (Recurrent Neural Network), LSTM (Long Short-Term Memory) for music generation. After thoroughly reviewing and understanding, out of these methods we chose CNN for emotion detection because of its high accuracy and LSTM for music generation as it can maintain information in memory for longer period of time. Dataset used for emotion detection was Fer2013 (the collection of 35,887 grayscale images of 48 X 48-pixel dimensions). The data consists of 48x48 pixel grayscale images of faces. The task was to categorize each face based on the emotion shown in the facial expression into one of seven categories (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral). Following are the statistics of the number of images present in the dataset for each emotion:

0: -4593 images- *Angry*
1: -547 images- *Disgust*

2: -5121 images- *Fear*
3: -8989 images- *Happy*
4: -6077 images- *Sad*
5: -4002 images- *Surprise*
6: -6198 images- *Neutral*

MIDI files were used to train the music generation model. We collected the MIDI files and prepared a dataset of music files for each emotion.

IV. METHODOLOGY

Mood Based Music player consists of two models Emotion detection and music generation model.

Emotion Detection

Emotion detection model uses Convolutional Neural Network (CNN) for recognizing facial expressions. In our proposed system, the CNN model is trained on FER2013(Facial Expression Recognition) dataset. The CNN model is trained with images as batch which contains 64 images for 20 epochs. The model will generate output with 7 possibilities of the input image. The weights are optimized using Adam optimization algorithm i.e. the network weights are updated in an iterative approach. Adam optimization technique is chosen over Stochastic gradient descent (SGD) because in Adam optimization the learning rates are learned on a per parameter basis unlike SGD which has a single global learning rate that is applied to all the parameters. L2 regularization technique is used to avoid overfitting of the model. Loss is calculated using categorical cross entropy which is usually used to calculate loss in a model which performs classification based on labels. In the system the webcam captures the image and then the captured image is fed as an input image to the model for detecting the emotion of the user. Using Haar-Cascade face detection algorithm, the face is detected in the captured image. The model now predicts the facial expression of the detected face.

Music Generation

The dataset used for music generation consists of music files of MIDI (Musical Instrument Digital Interface) format for each emotion. Unlike mp3, midi files cannot be played, as midi files do not contain actual audio. It contains information about the notes, chords, velocity of the notes, etc. Basically, they are an instructional file that provide information like what notes are being played, for how long are they played, how loud the notes are, etc. For music generation midi files are preferred over wav or other format files because comparatively they are small in size and very informative. Before training the model, these files are preprocessed i.e. encoded in the appropriate format that is more suitable for training the data. LSTM (Long Short-Term Memory) is used for music generation. Music generation requires a neural network which would remember a sequence for a longer period of time and RNN (Recurrent Neural Network) are capable of doing this. RNN consider each and every past event and the present event for predicting the future event. But in music generation, it is not necessary for remembering all the past events which is where LSTM play an important role in music generation. The optimization technique used in music generation is the same as the one used in emotion detection model. The loss function used during training is sparse cross entropy loss function. For generating music, the model is given an array consisting of randomly selected 50 notes. Now the model generates an array containing 300 notes by predicting the further notes. These notes are then decoded back to the original midi format. After decoding the midi file is then converted to mp3 format which is then played on the website after the emotion is detected.

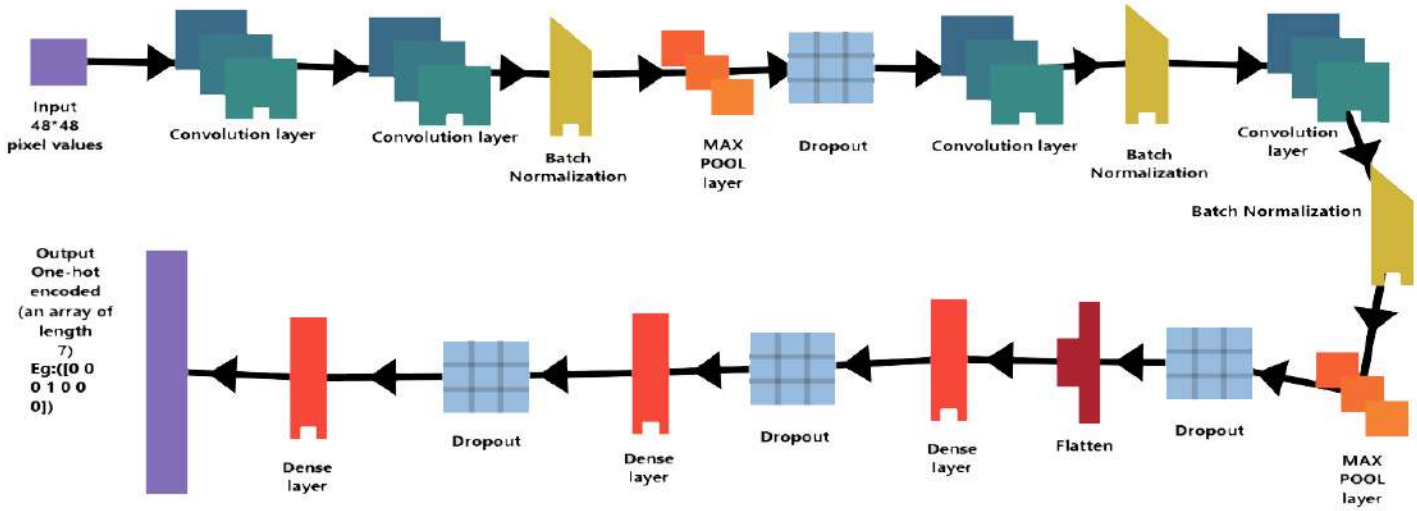


Fig 1: Emotion Detection Architecture

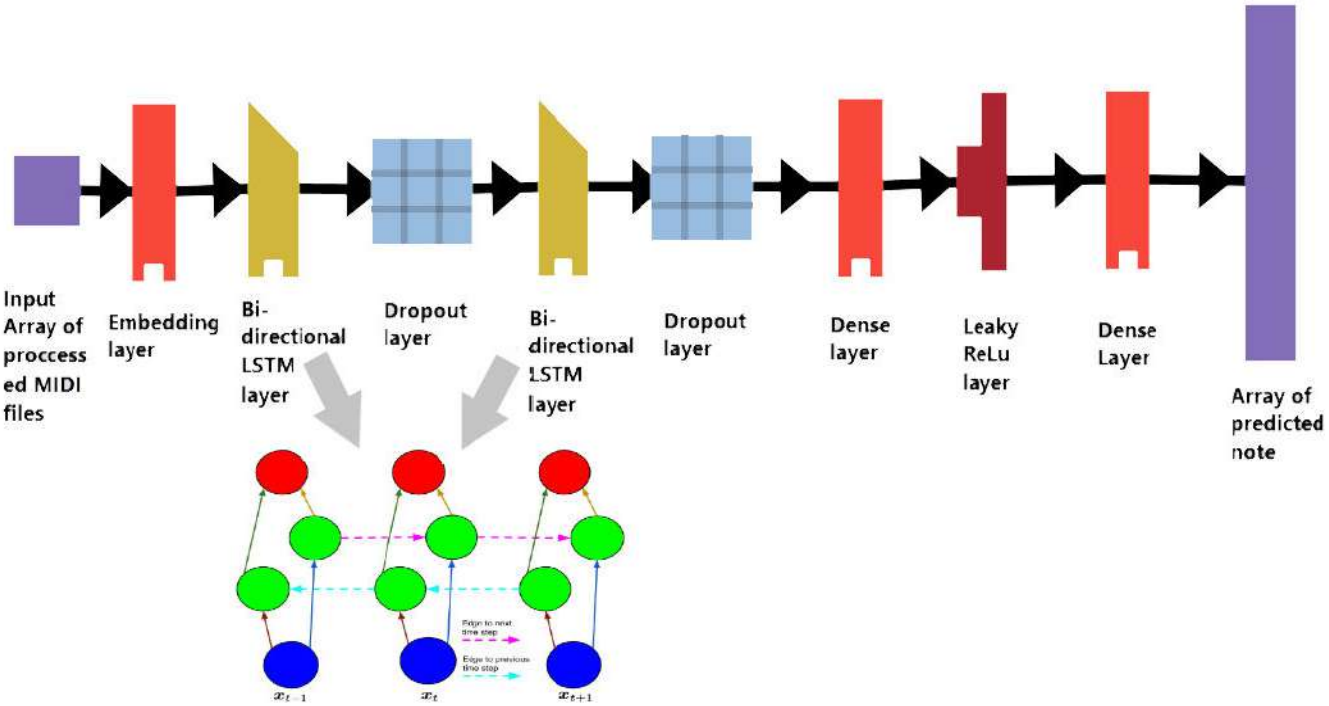


Fig 2: Music Generation Architecture

V. EXPERIMENTAL RESULT AND ANALYSIS

Tools Used

In Emotion Detection, Keras was primarily used for classification. Along with it the libraries used were Pandas, NumPy, OpenCV. For Music Generation, TensorFlow was used. PrettyMidi library was used for encoding and decoding of the midi files.

Results

The accuracy of the trained model is evaluated using the testing dataset which is split from the original dataset. 10% of the original dataset is used for testing and 10% for validation. The Emotion Detection model's accuracy is measured by directly comparing the model's predicted output to the actual output of the image. In this system, the model's accuracy is evaluated using Categorical Cross Entropy. Performance of a classification model in which output is a probability value between 0 and 1 is given by Cross-entropy. The results of the proposed application are shown below. The image is given as an input through webcam to the website. OpenCV is used to process the images taken by the webcam. The Fig shows emotions detected such as happy, sad, angry, surprise, neutral, fear images. The predicted array will show the probabilities of all the emotions. These probabilities lie between the range 0 and 1. The emotion which has the highest probability is taken as the final emotion of the user. The total accuracy of the emotion detection model is 64.6% for testing data. The validation accuracy is 69.7%. Given below is the demo of emotion detection model-

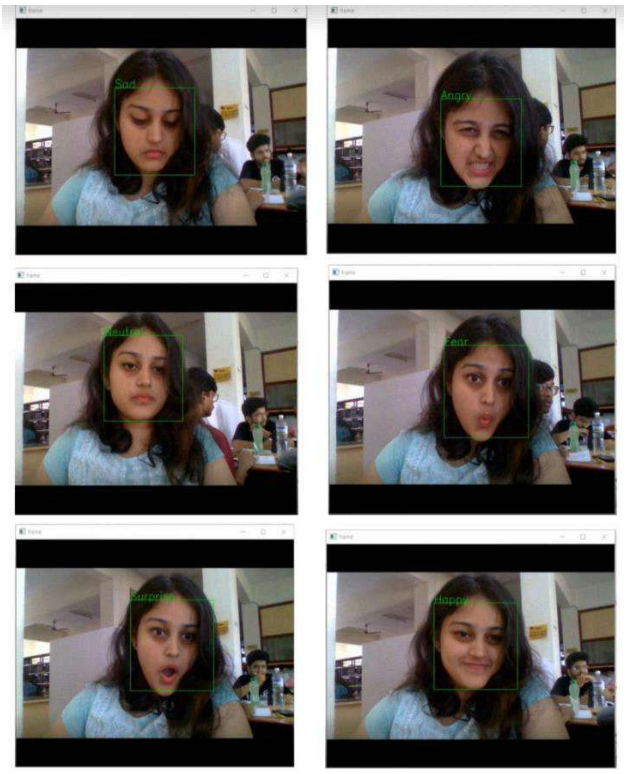


Fig 3: Output images with emotion detected

VI. FUTURE SCOPE

In future, this project can be further extended by using Voice Sentiment analysis can be applied for recognizing user's emotion along with the facial emotion recognition. In future mood-based music player can be incorporated in social media apps. It can also be used in several meditational apps for soothing user's mood.

VII. CONCLUSION

This mood-based music player detects the user's emotions and generates music according to the user's mood. In many existing systems, predefined playlists are provided to the user but in our proposed system the user is given a choice if he/she wants to listen to the generated tunes or he/she wants to listen to the predefined playlists. In this system, CNN works efficiently in recognizing the emotion of the user with an accuracy of 64.6%. The music generation model every time generates a unique tune for each emotion.

ACKNOWLEDGEMENT

We would like to express our gratitude towards our guide Prof. Gayatri Hegde who encourages us as well as helps us to solve our queries. We would like to thank her for her kind cooperation and encouragement which helped us in completion of this project. We give a special gratitude to our honorable principal Dr. Sandeep Joshi who always encourage us and motivate us to do innovative things that will increase our knowledge. We would also like to thank our H.O.D of Information Technology Department Dr. Satishkumar Varma for his guidance and also for giving us the opportunity to implement the project.

REFERENCES

[1] Abhishek R. Patel, Anusha Vollal, Pradnyesh B. Kadam, Shikha Yadav, Rahul M. Samant, "MoodyPlayer: A Mood based Music Player", International Journal of Computer Applications (0975 – 8887): Volume 141 – No.4, May 2016

[2] Rishi Madhok, Shivali Goel and Shweta Garg, "SentiMozart: Music Generation based on Emotions," Delhi Technological University, New Delhi, India: In Proceedings of the 10th International Conference on Agents and Artificial Intelligence (ICAART 2018)

[3] Maruthi Raja S K, Kumaran V, Keerthi Vasan A, Kavitha N, "Real Time Intelligent Emotional Music Player" Saranathan College of Engineering. Trichy, India: Journal for Research | Volume 03| Issue 01 | March 2017

[4] Karan Mistry, Prince Pathak , Prof. Suvarna Aranjo, ''Mood based Music Player" International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056 Volume: 04 Issue: 3 | Mar -2017

[5] Celina Jenefer C, Leena. S, Nirmala Devi M, Dr. J. SelvaKumar, ''A Survey: Expression Based Music Player" Department of Computer Science and Engineering, Sri Ramakrishna Engineering College, Coimbatore, Tamil Nadu, India:2017 IJSRST | Volume 3 | Issue 3

[6] Aditya Gupte, Arjun Naganarayanan, Manish Krishnan, ''Emotion Based Music Player – XBeats" Department of Computer Engineering, SIES GST, Mumbai University, Navi Mumbai, India: International Journal of Advanced Engineering Research and Science (IJAERS) [Vol-3, Issue-9, Sept- 2016]

[7] Hemanth P, Adarsh, Aswani C.B, Ajith P, Veena A Kumar, "EMO PLAYER: Emotion Based Music Player", Department of Computer Science, Saintgits College of Engineering, Pathamuttom, Kottayam, Kerala, International Research Journal of Engineering and Technology (IRJET) Volume: 05 Issue: 04 | Apr-2018

[8] Sri Charan Nimmagadda, "Emotion Based Music Player", Department of Computing Sciences Texas A&M University - Corpus Christi Corpus Christi, Texas

[9] Rahul Hirve, Shrigurudev Jagdale, Rushabh Banthia, Hilesh Kalal & K.R. Pathak, "EmoPlayer: An Emotion Based Music Player", Imperial Journal of Interdisciplinary Research (IJIR) Vol-2, Issue-5, 2016

[10] Harshada Sonkamble, Prof. Ujwala.V. Gaikwad, "Emotion Recognition Based on Efficient Self Organized Map", Terna Engineering College, Nerul, International Research Journal of Engineering and Technology (IRJET) Volume: 04 Issue: 04 | Apr -2017

Predicting Employees Performance using Data Mining Techniques

Samruddhi Gavas, Darshan Oswal, Ronish Rathod and Dr Madhu Nashipudimath

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

Abstract— Employee is the key element of the organization. The success or failure of an organization depends on employee performance. Human Resources Management (HRM) has become one of the essential interests of managers and decision-makers in almost all types of businesses to adopt plans for correctly discovering highly qualified employees. Accordingly, managers become interested in the employee. From here, the interest in the data mining (DM) role has been growing and its objective is the discovery of knowledge from huge amounts of data. Data Mining techniques were utilized to build a classification model for predicting the performance of employees using a real dataset. There are various data classification techniques such as DT, Support Vector Machine (SVM), Naïve Bayes classifier, J48, KNN and others. The major goal of this project is to predict the employee's performance using machine learning algorithms for data mining. For this, we use different methods such as Support Vector Machine, Logistic Regression and Neural Network. This System rates the performance of employees and gives a predictive analysis of data mining techniques.

Keywords—Neural Network, Logistic Regression, SVM, Classification, Data Mining, Employees Performance Rating

1. Introduction

Human Resource Management(HRM) has a leading role in deciding the competitiveness and effectiveness for better continuation. It becomes the responsibility of the HRM to allocate the best employees to the appropriate job at the right time, train and qualify them, and build evaluation systems to monitor their performance and an attempt to preserve the potential talents of employees. With the advancement and growth of technologies in business organizations, HR employees need not handle the massive amount of data manually any further. This data is very important for the decision-makers, but there is a challenge to mine and get the best and useful data from these huge data. From here, the role of Data Mining comes.

Data Mining is considered as a recently emerging analysis and predictive tool, because of the existence and multiplicity of massive amounts of data containing huge hidden unknown knowledge. Knowledge can be extracted through various methods and one of them is by using the Data Mining technique. Data Mining techniques provide an approach to utilize different Data Mining tasks such as classification, association, and clustering used to extract hidden knowledge from huge amounts of data. With classification, Predictive models have the specific target of enabling us to predict the unknown values of variables depending on interest previously known values of other variables. Data Mining is the next big revolutionary field that is redefining the industry, be it in terms of technology or research.

Machine Learning is an application of artificial intelligence which allows the machine to learn from examples and experience, and all that without being explicitly programmed. So instead of writing the code, we need to feed data to the generic algorithm, and the algorithm/ machine builds the logic based on the given data. The Machine learning algorithm has increasing computational power. So it helps to discover new knowledge from large databases. That knowledge is very useful for business analysis.

Classification techniques are supervised learning techniques that classify data items into predefined class labels. It is one of the most useful techniques in data mining to build classification models from an input data set. The used techniques commonly build models that are used to predict future data trends. There are various data classification techniques such as DT, Support Vector Machine, Naïve Bayes classifier, and others. With classification, the generated model will be able to predict a class for given data depending on previously learned information from historical data. In this project, the classification process is executed through Support Vector Machine, Logistic Regression and Neural Network on various attributes. We use the Feature Selection algorithm for extracting useful

information from the dataset. The main objectives of the present study were extracted to performance based on a real dataset to get real and support the decision-makers discover potential talents of employees. The objectives are as follows :

1. To study the data mining techniques and identify their limitations.
2. Evaluate the performance of employees.
3. To understand various data classification techniques such as Logistic Regression, Support Vector Machine and Neural Network.
4. To support the decision-makers in gathering a dataset of predictive variables.
5. Identification of different factors which affects employees behaviour and performance.
6. Using proposed DM classification techniques for constructing a predictive model and identifying relationships between the most important factors affecting the whole efficiency of the model.

The report is organized into five chapters: The introduction is given in Chapter 1. It describes the fundamental terms used in this project. It motivates me to study and understand the different techniques used in this work. This chapter also presents an outline of the objective of the report. Chapter 2 describes the review of the relevant various techniques in the literature systems. It describes the pros and cons of each technique. Chapter 3 explores the Existing System Architecture and Proposed System Architecture model for constructing the proposed model. It describes the major approaches used in this work. It also describes the software and hardware requirements for the project and the different algorithms used. The Evaluation Parameters i.e input and output are mentioned in Chapter 4. The summary and future research directions of the report is presented in Chapter 5.

2. Literature Survey

Several studies used data mining for extracting rules and predicting certain behaviors in several areas of science, information technology, human resources, education, biology and medicine. For example, Dr. J. Krishna, D. Deekshitha, P. Neelaveni, S.Leelavathi, V.Lakshmi Sai (2020) and Mona Nasr ,Essam Shaaban, Ahmed Sami in different locations to significant results for supporting the HR executives and the decision makers. There have been three classification techniques that are classifier SVM, DT, and Naïve Bayes.

This study concluded that SVM was found to be the most appropriate classifier for the construction of the predictive model, where it had the highest predictive accuracy through all three tests with the highest percentage of 86.90%.[1,2].

Zarmina Jafar, Dr. Waheed Noor, Zartash Kanwal (2019) compare different feature selection methods Naive Bayes, Logistic Regression, and J48 classifier Existing feature selection algorithms may not be able to generate a valid subset of features for the classification of many different regions. Although some algorithms may reduce features, their classification accuracy is not high. The proposed information selection algorithm based on conditional equivalence produces high efficiency and small characteristics in several different data sets, and the classification accuracy is higher [3].

Farhad Sheybani (2019) discussed the Job satisfaction and and determining the factors affecting it using the CHAID Decision Tree Data Mining Algorithm. At first, the data mining tool was used to prepare this data and then, using the CHAID decision tree in Clementine 12.0 software, factors affecting job satisfaction were investigated in the sample individuals. According to the results, it concluded that the lack of individuals' pride to work for their employer can have a great impact on the full job dissatisfaction [4].

Rahul yedida, Rakshit Vahi, Abhilash, Rahul Reddy, Rahul J, Deepti Kulkarni (2018) discusses the method of predicting whether an employee of a company will leave or not, using the k-Nearest Neighbors algorithm. We use evaluation of employee performance, average monthly hours at work and number of years spent in the company, among others, as our features. Other approaches to this problem include the use of ANNs, decision trees and logistic regression. The dataset was split, using 70% for training the algorithm and 30% for (2019) presented a study for predicting the employees'

testing it, achieving an accuracy of 94.32%. attrition [5].

Farhad Sheybani (2019) found CHAID Decision Tre Data Mining Algorithm suitable for the dataset[4]. Zarmina Jafar, Dr. Waheed Noor, Zartash Kanwal (2019) concluded J48 to be the suitable algorithm among Naive Bayes, Logistic Regression, Bayes Network and OneR [3]. Rahul yedida, Rakshit Vahi, Abhilash, Rahul Reddy, Rahul J, Deepti Kulkarni predicted k-Nearest Neighbors algorithm is accurate amongst ANNs, decision trees and logistic regression [5] .Dr. J. Krishna, D. Deekshitha, P. Neelaveni, S.Leelavathi, V.Lakshmi Sai (2020) and Mona Nasr, Essam Shaaban, Ahmed Sami (2019) predicted that SVM classification technique is the highest predictive accuracy technique among the classifier SVM, DT, and Naïve Bayes[1,2].

In General, this paper is an initiative to investigate Data Mining tasks, especially classification tasks, for supporting decision makers and by studying the main factors of their employees that may positively affect their performance. The paper applied some of the classification techniques to build a proposed model for supporting the prediction of the employees' performance. This study attempts to use classification techniques in data mining to determine the employee's performance by predicting their performance based on the past experience knowledge from employee databases.

2.1 Summary of Related Work

The summary of methods used in literature is given in Table 1.

Table 1 Summary of literature survey

Literature	Method 1	Method 2	Hybrid
Dr. J. Krishna, D. Deekshitha, P. Neelaveni, S.Leelavathi, V.Lakshmi Sai 2020 [1]	Yes	No	No
Mona Nasr EssamShaaban,Ahmed Sami 2019 [2]	No	Yes	No
Farhad Sheybani 2019 [3]	Yes	Yes	No
Zarmina Jaffar, Dr.Waheed Noor, Zartash Kanwal 2019 [4]	Yes	Yes	Yes
Rahul yedida, Rakshit Vahi, Abhilash, Rahul Reddy, Deepti Kulkarni2018[5]	Yes	Yes	No

3. Proposed Work

Controlling the computer mouse using the eyes movement requires a fast and effective algorithm, that's brought us to decrease the running time of the tool to the minimum by dividing the operation into few steps and using a tracking algorithm in order to avoid unnecessary calculations.

3.1 SystemArchitecture

The system architecture is given in Figure 1. Each block is described in this Section.

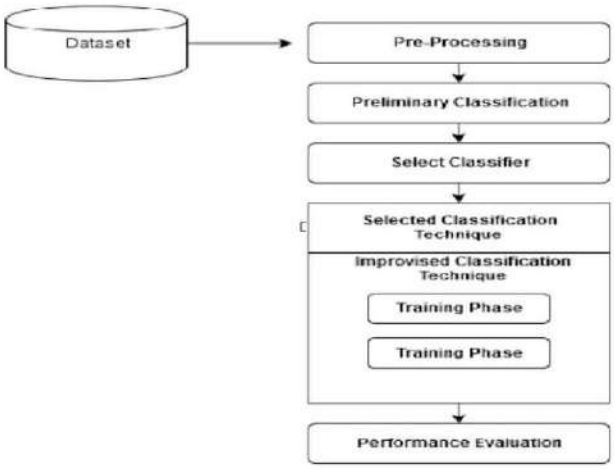


Fig. 1 Proposed system architecture

A. Data Collection and Understanding Process: The idea study is building a classification model for predicting the employees' performance based on a real dataset to get real and significant results for supporting the HR executives and the decision makers.Data set is a collection of data. Most commonly a data set corresponds to the contents of a single database, where every column of the table represents a particular variable, and each row corresponds to a member of the dataset. For our project we take employee data which contains 1200 records and 28 fields including categorical and numeric features. Each record in the employee data set represents a single employee information and each field in the record represents a feature of that particular employee.

B. Data Preparation and Pre-processing: After the process of data collection finished, the process of preparing the data is performed, the raw data contained instances that were not applicable.After the process of data collection finished, the process of preparing the data is performed, the raw data contained instances that were not applicable. It is important to refine this data so that it can be suitable for the models and generate better results. Once the data is selected, the third phase is to prepare this data. This phase includes tasks like cleaning, transformation and removing the unwanted data. The data of the employee had various attributes which were not relevant, i.e. was not giving any useful information, like Employee Number, Employee count, etc., hence these attributes were removed in the process of data cleaning.

C. Feature Selection: Feature selection is a one of the main concepts of DM and Machine Learning. Where, it is a process of selecting necessary useful variables in a dataset to improve the results of machine learning and make it more accurate.

There are a lot of columns in the predictor variable. So,

the correlation coefficient is calculated to see which of their robust initial training and then from ongoing them methods. From there, we also get the top factors which affect performance. We can see that the most important features selected were Department, Job Role, streams in favour of those most likely to be accurate. Environment Satisfaction, Last Salary Hike Percent, Work That means a preference is put on the input streams that Life Balance, Experience Years At This Company, have a higher weight; and the higher the weight, the Experience Years In Current Role, Years Since Last more influence that unit has on another. The process of Promotion, Years With Current and Label Encoding was also used for feature of this transformation.

D. Classification Model: Data classification is the process of organizing data into categories for its most effective and efficient use. There have been three classification techniques that are classifier SVM, Logistic Regression, and Neural Network. Such classification techniques applied to the data set to construct the performance prediction model of the employees in order to obtain the most suitable DM methodology and the most efficient variables that can influence and forecast the salary of the employees.

Logistic Regression is a mathematical model used in statistics to estimate the probability of an event occurring using some previous data. Logistic Regression works with binary data, where either the event happens (1) or the event does not happen (0). Logistic regression is generally used for classification purposes. When the number of possible outcomes is only two it is called Binary logistic Regression.

SVM is considered as one of the most effective supervised machine learning techniques that has a straightforward structure and high ability for classification. Moreover, SVM is recognized as the appropriate technique in DM for classification particularly on both linear and non-linear decision margins where high accuracy of model can be produced. SVM has many advantages such as it has no ceiling on the number of attributes and depends on the kernel trick for building the model through expert knowledge on the problem via kernel adjustment. Sequential Minimal Optimization (SMO) is a SVM algorithm.

Neural networks modify themselves as they learn from

are important and these are then used for training self-learning that they experience by processing additional information. A simple learning model applied by neural networks is the process of weighting input Manager. These were reducing predictable errors through weight, is done selected because their correlation coefficient with through gradient descent algorithms. Finally, output Performance Rating was more than 0.1. Standardization units are the end part of the process; this is where the

network responds to the data that was put in initially, and can now be processed.

E. Evaluation: Employee evaluations are an important part of maintaining a motivated and skilled workforce. Every company maintains a confidential report form for measuring the quality of an employee throughout the year. The rating scale is the user input to the organization.

Employees Performance Result	Employees Performance Rating
Good	2
Very Good	3
Excellent	4

Table 2 Performance Rating Scale

Top 3 factors which affect the employee performance are

1. Employee EnvironmentSatisfaction
2. Employee Last Salary Hike Percent
3. Years Since Last Promotion

Graphical representation for performance of each department is shown below in Fig 2

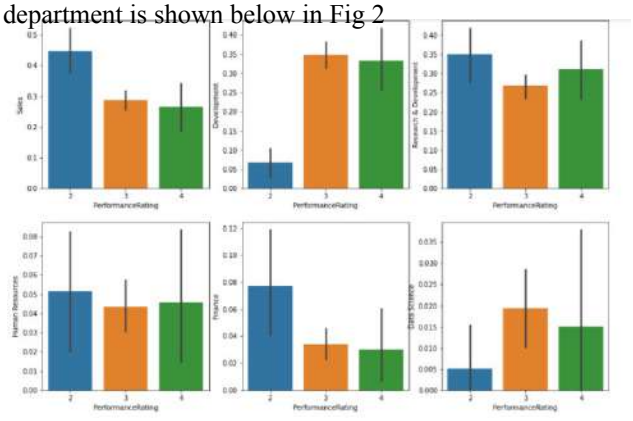


Fig. 2 bar graph for performance of each department

E. Output Block Description: Firstly HR can login to this system by entering the correct username and

Fig 3 GUI for HR Login

In the section below, we used algorithms like Logistic Regression, Support Vector Machine and Artificial Neural Network to calculate the accuracy and found out that the Artificial Neural Network gives the maximum accuracy of
0.87%.

Fig. 4Accuracy Prediction Result

Live prediction window displays the prediction result on all three techniques. The prediction result is given in the form of rating as shown in Fig 5.

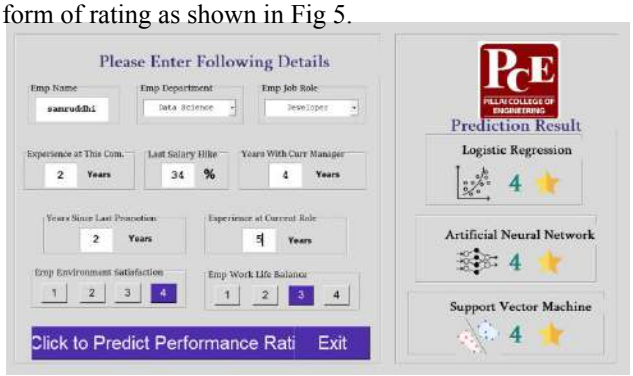


Fig.5 Live PredictionAnalysis

password. After login HR can select between Algorithm Comparison
and Live Prediction as shown in fig 3.

Algorithms is shown below in the table . According to the table , the results indicated that the Neural Network technique has the highest prediction accuracy through using the most effective factors.

3 RequirementAnalysis

The implementation detail is given in this section.

3.1 Software

The software requirements of the system are described below. The operating systems used will be windows 7& above. Programming languages used are Python, HTML5, CSS3, Bootstrap.

3.2 Hardware

The hardware required by the system to be developed is given below. The main memory required is 8 GB & above so that the whole program can reside on the same memory at once. This will avoid the requirement to swap the memory contents of the system. Hard disk drive is required to store the program permanently on the storage so that the loss of power will not affect the availability of the program. Processor is required to process the data quickly on the system. A Computer/Laptop is required to enable the user to interact with the system while on the go.

3.3 Dataset and Parameters

Data set is a collection of data. Most commonly a data set corresponds to the contents of a single database, where every column of the table represents a particular variable, and each row corresponds to a member of the dataset. For our project we take employee data which contains 1200 records and 28 fields including categorical and numeric correlation matrix There are a lot of columns in the predictor variable. So, the correlation coefficient is calculated to see which of them are important and these are then used for training methods. From there, we also get the top factors which affect performance. We can see the most important features selected. These were selected because their correlation coefficient with Performance Rating was more than 0.1.

The Accuracy Percentages table for Prediction features. Each record in the employee data set represents a single employee information and each field in the record represents a feature of that particular employee.

A statistical test to check if the attributes have any correlation among each other was done with the

Table 4 FinalAttributes used in the module

Sr. No	Attributes
1	EmpDepartment
2	EmpJobRole
3	EmpEnvironmentSatisfaction
4	EmpLastSalaryHikePercent
5	EmpWorkLifeBalance
6	ExperienceYearsAtThisCompany
7	ExperienceYearsInCurrentRole
8	YearsSinceLastPromotion
9	YearsWithCurrManager
10	PerformanceRating

ACKNOWLEDGMENT

It is our privilege to express our sincerest regards to our supervisor Prof. Madhu Nashipudimath 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 to our Head Department Dr. Satishkumar Varma
and our Principal Dr. Sandeep M. Joshi for encouraging us .

.

References

1. *Dr. J. Krishna, D. Deekshitha, P. Neelaveni, S.Leelavathi, V.Lakshmi Sai,* (2020), "Employee Performance Predicting Using Classification Techniques", C.S.E, Annamacharya Institute of Technology,India.
2. *Mona Nasr, Essam Shaaban ,Ahmed Samir* "A proposed Model for Predicting Employees' Performance Using Data Mining Techniques: Egyptian Case Study"International Journal of Computer Science and Information Security (IJCSIS),Vol. 17, No. 1, January 2019.
3. *Farhad Sheybani*, (2019), "Predicting the Individuals' job satisfaction and determining the factors affecting it using the CHAID Decision Tree Data Mining Algorithm", The National Opinion Research Center of the United States.
4. *Zarmina Jaffar, Dr. Waheed Noor, Zartash Kanwal*, (2019), "Predictive Human Resource Analytics Using Data Mining Classification Techniques", University of Balochistan Quetta, Pakistan.
5. *Rahul yedida, Rakshit Vahi, Abhilash, Rahul Reddy, Rahul J, Deepti Kulkarni,* (2018), "Employee Attrition Prediction", PESIT-BSC Bangalore.
6. *R Shiva Shankar, J Rajanikanth, V.V.Siva Rama Raju, K VSSR Murthy*, "Prediction of Employees attrition using data mining" SRKR Engineering College, Bhimavaram, India
7. *Hamidah Jantan, Maridah Puteh, Abdul Razak Hamdan, Zulaiha Amli Othman,* "Applying Data Mining Classification Techniques for Employee's Performance Prediction" Universiti Teknologi MARA(UiTM) Terenggan
Software-Industry" Jain University, Bangalore.
8. *Gaurav Singh Thakur, Anubhav Gupta, Sangita*
9. *Hamidah Jantan, Abdul Razak Hamdan, and Zulaiha Ali Othman,* "Knowledge Discovery Techniques for Talent Forecasting in Human Resource Application", World Academy of Science, Engineering and Technology International Journal of Industrial and Manufacturing Engineering Vol:3, No:2, 2009
10. *Qasem A. Al-Radaideh, Eman Al Nagi*, "Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance" Vol. 3, No. 2, 2012
11. *Tejas Raut, Priya Kale, Rashmi Sonkusare, A. K. Gaikwad*, "Employee Performance Prediction System using Data Mining", Volume: 07 Issue: 02 | Feb 2020
12. *S. E. Viswapriya*, "Survey on Predicting Performance of An Employee using Data Mining Techniques", Vol. 8 Issue 10, October-2019
13. *Komal Vikas Kaware, Pratiksha Maruti Viveki, Shrutika Dnyanesh Lokhande, Nikita Neminath Avadhut , Vaishnavi Somnath Pattanshetti , S. A. Shinde,* "A Survey on Predicting Employee's Performance", Journal of Network Security and Data Mining Volume 2 Issue 3
14. *Ananya Sarker, S.M. Shamim, Dr. Md. Shahidul Zama & Md. Mustafizur Rahman*, "Employee's Performance Analysis and Prediction using K-Means Clustering & Decision Tree Algorithm", Volume 18 Issue 1 Version 1.0 Year 2018
15. *N. Magesh M.E., Dr. P. Thangaraj Ph.D, S. Sivagobika, S. Praba, R. Mohana Priya,* "Evaluating The Performance Of An Employee Using Decision Tree Algorithm", Vol. 2 Issue 4, April - 2013