Journal of
Information Technology

Volume 8, Issue 1, 2020-21

JIT
Volume 8
Issue 1
2020-21

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 8, Issue 1, 2020-21

Message

Dr. Sandeep M. Joshi
Principal, PCE

Learning is the process of constantly updating in an ever changing word. Project based learning is a model where both the teacher and the student try to understand each other and create a congenial ambience and easy adaptable teamwork which help them to achieve academic excellence and empower them to be lifelong learners.

I take this opportunity to thank parents, the students and teachers for having their faith and confidence in our Project Based model of learning and bringing out this issue as an outcome. Your faith in us is our driving force.

My best wishes to the teachers, students and lab assistants for a successful endeavour.

Editorial

Dr. Satishkumar L Varma
Editor-in -Chief

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

It gives me immense pleasure to bring our department journal during pandemic Covid-19 by working from home. Our teachers and students have mediculously planned and worked on wrting and submiting technical papers.

This journal focuses on a variety of topics such as Deep Learning, Natural Language Processing, Machine Learning, IoT and Security. This issue of PCE JIT contains the application of various technologies such as Deep Learning for Number Plate Detection of cars and Fake Profile Detection. It also discusses the security applications such as Penetration Testing, IoT applications such as Soil Fertilization System.

This issue covers ten 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 Deep Learning, Natural Language Processing, Machine Learning, IoT and Security.

I wish a successful and fruitful publication life with our department journals.

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

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 32 refereed Journal papers and more than 36 papers in referred National as well as International Conferences including IEEE, Springer and IET with a second best paper award at National level paper presentation competition in Threshold–2000. He is recognized as Teacher of University of Mumbai in Ph.D Degree in Computer Engineering and Information Technology. His delivered talks include Image Processing, Object Oriented Analysis and Design, MATLAB, Scilab, Hadoop, LaTeX, Android, Python, R, Google Scripts and Docs. He is a member of Technical Professional society in IEEE, ISTE, and CSI. His research interests involve Digital Image and Video Processing, Medical Imaging, AI and Machine Learning, Soft Computing, Data Mining and Information Retrieval.

Sushopti Gawade 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 8 conference and journal purblications and has attended 10 FDP. Her area of interst includes Operating system, Cloud Computing, Big Data Analytics and Distributed Systems.

Ninad Gaikwad received his M.E degree in Information Technology from Engineering from Pillai College of Engineering, Mumbai University in 2020. He has completed his BE in Information Technology from Indira College of Engineering and Management, Savitribai Phule Pune University. He is Currently working as Assistant Professor in the Department of Information Technology, Pillai College of Engineering, New Panvel, Maharashtra. He has three International Conference publications. His area of interest include DevOps, Artificial Intelligence and Machine Learning.

AUTOMATIC NUMBER PLATE DETECTION AND RECOGNITION USING DEEP LEARNING

Mohammed Sameel Shaikh
Information Technology
Pillai College of Engineering Panvel, Navi Mumbai

Email:mohsa17ite@student.mes.ac.in

Yogitha Nilekani
Information Technology
Pillai College of Engineering Panvel, Navi Mumbai

Email:nilekaniyogya17ite@student.mes.ac.in

Harshada Shinde
Information Technology
Pillai College of Engineering Panvel, Navi Mumbai

Email:shindeharpr17ite@student.mes.ac.in

Satishkumar Varma
Information Technology
Pillai College of Engineering Panvel, Navi Mumbai

Email:vsat2k@mes.ac.in

Abstract: ***The treatment of vehicles is possible by detecting number plates. The system of vehicle number plate detection is done by extracting the text from the number plate by locating the alphanumeric characters on a number plate using image processing techniques. The reading of license plates requires an intelligent system and there are various methods to train the system to recognize characters. In this paper, vehicle image capturing, preprocessing of the image, detection of number plates in the image, segmentation of characters and recognition of characters on the number plate is implemented by using deep learning techniques by analyzing proper deep network structures. An automatic system is developed using Python to perform detection as well as recognition of a car number plate. The various classifiers such as Artificial Neural Networks/Deep Learning, Support Vector Machine are used to train with features extracted from a set of images by using the concept of Convolutional Neural Networks. The video data is taken from surveillance cameras to train and test the result of number plate detection and recognition systems. Such systems are used in the case of vehicle usage in illegal activities, invalid number plates, stolen cars, etc. It can also be used in highway electronic toll collection and speed detection.***

Keywords- **Convolutional Neural Networks, Optical Character Recognition, Deep Learning, Artificial Neural Network, Vehicle License Plate Recognition, Darknet.**

I. INTRODUCTION

Monitoring vehicles has become a tedious task because of the tremendous increase in the number of vehicles. There is an increasing demand for an automated vehicle identification system. Since many years there has been an increasing research interest among various researchers in this domain related to extracting text from video. One of its applications is locating the number plate in a video. Image processing techniques are used to efficiently recognise the number plate from a picture or a video. The applications of automatic number plate detection and recognition include identifying vehicles by their number plates for faster traffic management at parking areas, better security and prevention of car theft, the ability to automate access control systems, allowing new and more effective law enforcement.

As the population in India is growing massively, with the increase in the number of vehicles there is a pressing need to come up with a better and smart way to handle the increasing number of vehicles and shape the traffic efficiently. Whereas with the increasing crimes in the present scenario it is

equally important to generate a system that keeps a track of vehicles by reading their number plates. Machine learning has the potential to ease the whole process of analysing data effectively. In deep learning, a convolutional neural network (CNN) is a class of deep neural networks, which is mostly applied to analyzing visual imagery. The YOLO (You Only Look Once) model is used for detecting the number plate and the Tesseract is used for Optical Character Recognition (OCR). The video is first converted into a set of frames and from each frame, the number plate is detected and sent for recognition. In the recognition process, each character is segmented and recognised.

II. LITERATURE SURVEY

During the Literature survey, we collected information about various techniques used for license plate detection and recognition as shown in Table 1.

1. Far number plates detection

Anisha goyal [4] proposes the automatic number plate recognition system by using vehicle license number plates. The system uses image processing techniques for recognizing the vehicle from the database stored in the computer by the user. The system works for a wide variety of conditions and distinctive sorts of number plates. The system is executed in Matlab and performance is tried on genuine images.

2. Recognition System for Criminal Surveillance

Siddharth U. Mishra [9] proposes the automatic license plate recognition system using image processing which has a conversion of gray scale image, noise reduction, contrast enhancement using histogram equalization and plate localization. Mathematical morphology is used to detect the region. Using the Sobel edge detector the highlighted regions with a high edge magnitude and high edge variance are identified. From the input image based on the threshold value, the edge of the number plate gets detected. For character segmentation, the Matlab toolbox function provides a function called regionprops() is used. It measures a set of properties for each labelled region in the label matrix.

3. Recognition Using Random Forest Classifier

Zuhaib Akhtar and Rashid Ali [10] paper proposed an algorithm that has preprocessing, number plate localization, character segmentation and character recognition steps. Preprocessing includes converting RGB image to grayscale image, removing noise by using a bilateral filter, then increasing the contrast of the image using CLASH, converting the image to a binary image and finally dilating the image. By using Sobel vertical edge detection, number plate localization extracts the number plate region from the image. Character segmentation first removes the redundant portion of the number plate then segments the individual characters from the extracted number plate. Character recognition recognizes individual optical characters by using random forest classification algorithms.

4. Chirag Patel [3] paper proposes a comprehensive study of recent development and future trends in ANPR, which can be useful to researchers who are involved in such developments. It also states that certain factors like different illumination conditions, vehicle shadow and non-uniform size of license plate characters, background color and different fonts affect the performance of ANPR.

5. Recognition using OCR technique

In this paper [7] the accuracy of the OCR technique is checked and evaluated. Their proposed algorithm is based on Template matching. The algorithm crops the image to select characters, then the image undergoes a black and white transformation. After noise reduction, numbers and alphabets are compared with templates to get the result. They also found out that there are some factors that affect the effectiveness of template matching based on OCR technique i.e. font type, noise in image, tilting etc.

Table 1. Literature summary on automatic number plate detection and recognition system techniques.

Literature	Automatic Number Plate Detection & Recognition System Techniques				
	ANN	Tesseract OCR	Template Matching	ICR	CNN
Chirag Patel et al. 2013 [3]	Yes	Yes	Yes		
Er. Kavneet Kaur et al. 2013 [7]			Yes	Yes	
Siddharth U. Mishra et al. 2014 [9]					Yes
Mrs. J. V. Bagade, et al. [14]	Yes				
Anisha Goyal et al. 2016 [4]		Yes			Yes

III. SYSTEM METHODOLOGY

A. Methodology

Systems implementing a number plate detection and recognition extract the frames of images from the video obtained from surveillance cameras [1][2]. The YOLO model detects the number plate from the extracted frames. After detection, the number plates are used for recognition where the characters in the number plate are segmented and recognised separately using Convolutional Neural Network [5].

B. System Architecture

The proposed system architecture is given in Fig. 1. The architecture is described in this section with detailed explanation of each block associated with it.

Dataset Preparation: It is the initial step of the number plate detection and recognition. During the training process, the annotations of the open image dataset are converted into the format accessible by the YOLO model. The testing process includes processing the image by converting it into a set of frames from the video. The extracted frames are then sent for detection.

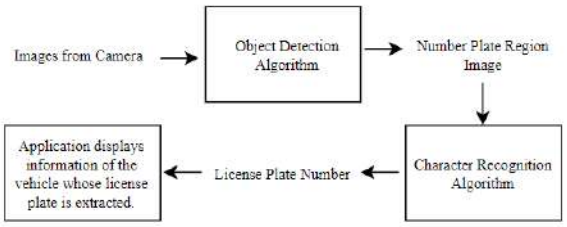


Fig. 1 System architecture for number plate detection and recognition

Licence plate detection: You Only Look Once or YOLO is one of the fastest real-time object detection algorithms (45 frames per seconds) as compared to the R-CNN family (R-CNN, Fast R-CNN, Faster R-CNN, etc.) For localising the objects in images, the R-CNN uses regions. The CNN model is applied to multiple regions and within the image high scoring regions are considered as object detected. But YOLO on the other hand selects some regions and applies a neural network to the entire image to predict bounding boxes and their probabilities. The YOLO method makes use of the Darknet-53 Model architecture as seen in Fig. 2, which has incorporated multiple CNN models as layers to enhance the system for accurate detection.

Convolutional Neural Network: Convolutional Neural Network, a type of Artificial Neural Network is used in image processing or recognition as it deals with visual imagery. The CNN model forms the heart of license plate detection as the YOLO models used are an intelligent version of CNN model which divides the images into different regions using a single neural network and does computations to predict bounding boxes within the image and the probabilities of these specific regions can be used to determine the actual bounding box in the given input image.

Yolov3: YOLOv3 uses a variant of Darknet architecture, that originally contains a 53 layer network trained on Imagenet, as shown in Fig. 2. For the detection task, in addition to 53 layers, 53 additional layers are stacked onto it, which gives us a 106 layer fully convolutional architecture for YOLOv3. The implemented result of YOLOv3 stating the accuracy of detection of the number plates is shown in Fig. 4.

Yolov4: YOLOv4 [13] architecture comprises CSPDarknet53(backbone), spatial, PANet path-aggregation(neck), pyramid pooling module and YOLOv3(head). The latest version of mAP (accuracy) and FPS (frame rate per second) are improved by 10% and 12%, respectively with respect to YOLOv3. An additional Convolutional Neural Network (CNN) layer has been added in YOLOv4 [12]. The implemented result of YOLOv4 stating the accuracy of detection of the number plates is shown in Fig. 5.

Licence plate recognition: The recognition of characters [6] of the vehicle number plate is done by OCR [8] techniques. OCR helps in recognizing text inside images, such as photos and scanned documents. OCR technology is mainly used to convert written text (typed, handwritten or printed) from any kind of images into machine-readable text data.

	Type	Filters	Size	Output
	Convolutional	32	3 x 3	256 x 256
	Convolutional	64	3 x 3/2	128 x 128
1x	Convolutional	32	1 x 1	
	Convolutional	64	3 x 3	
	Residual			128 x 128
	Convolutional	128	3 x 3/2	64 x 64
2x	Convolutional	64	1 x 1	
	Convolutional	128	3 x 3	
	Residual			64 x 64
	Convolutional	256	3 x 3/2	32 x 32
8x	Convolutional	128	1 x 1	
	Convolutional	256	3 x 3	
	Residual			32 x 32
	Convolutional	512	3 x 3/2	16 x 16
8x	Convolutional	256	1 x 1	
	Convolutional	512	3 x 3	
	Residual			16 x 16
	Convolutional	1024	3 x 3/2	8 x 8
4x	Convolutional	512	1 x 1	
	Convolutional	1024	3 x 3	
	Residual			8 x 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

Fig. 2 Darknet 53 Model [11]

Image Processing: The image obtained from the object detection model goes through certain image processing techniques to allow accurate recognition of license number.. The image processing techniques applied are: Grayscale conversion, Thresholding and Gaussian Blur. **Grayscale conversion** of the image is done to enhance speed, as grayscale conversion allows for faster recognition and color information does not help in recognition of characters. Once the image is converted to grayscale, **thresholding** is done on the image to eliminate noise, thus helping in increasing accuracy of the OCR model. Finally, a **Gaussian Blur** filter is applied as in some scenarios, noise is inserted into an image when image binarization is done. The image obtained after applying image processing techniques is then passed to Tesseract for recognition of alphanumeric characters.

Tesseract: Tesseract is a qualitative OCR technique for recognition of number plates. It uses a two-step approach that is called adaptive recognition. It requires one data stage for character recognition, then the second stage to fulfil any letters by letters that matches the word or sentence context.

IV. REQUIREMENT ANALYSIS

The implementation details are given in this section.

A. Software

The system requires preprocessing and in-depth study of the data for detection which is possible in the YOLO model. The project also requires GPU based computations which is possible by using tensorflow-gpu module. The technologies used are python, OpenCV, and Tensorflow.

B. Hardware

To deal with lots of data and the continuous nature of the environment one requires at least 4GB of RAM. A surveillance camera to continuously monitor the vehicles passing by and pass the real-time fetched environment to the system.

C. Dataset and Parameters

Standard datasets are taken from the Vehicle Registration Plate dataset Google Open Image Dataset. In the scope of this project, 1000 images are used. 850 images are used for training YOLO models and 150 images are used for testing the model. The given dataset consists of 1000 images of Vehicle Registration Plates with the annotations of the bounding boxes on these images which is used for training and testing purposes.

V. EVALUATION METRICS

mAP(mean average precision) is the average of AP. The mean average precision compares the ground-truth bounding box with respect to the detected box and returns a score. Higher score means that the model is more accurate in its detection.

VI. RESULTS

The result is described in the terms of the accuracy of our model yolov3 and yolov4 in Fig. 3.

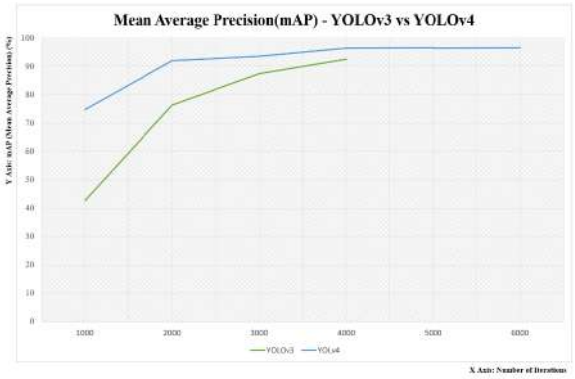


Fig. 3. Mean Average Precision (mAP) - YOLOv3 vs YOLOv4

The mAP (mean average precision) is calculated by the system and used for result and evaluation.

Table 2. Performance Evaluation of YOLOv3 and YOLOv4 techniques

Methods	**Image Size**	**Max batches (2000* # classes)**	**Filters (# class + 5)*3**	**mAP**
YOLO v3 Object Detection Algorithm	416 x 416	4000 (min - 4000)	18	92.5
YOLO v4 Object Detection Algorithm	416 x 416	6000 (min - 6000)	18	97.73

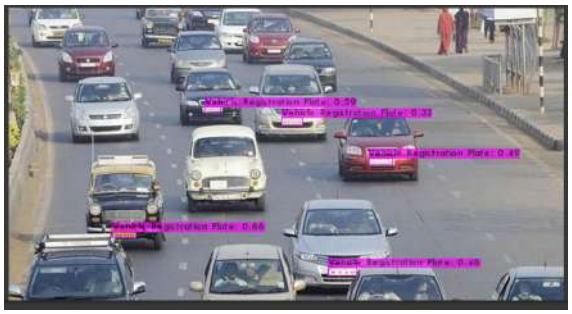


Fig. 4 Result showing the labelling of Number Plates for vehicles after training and detection using YOLOv3 technique over Open Image Dataset.

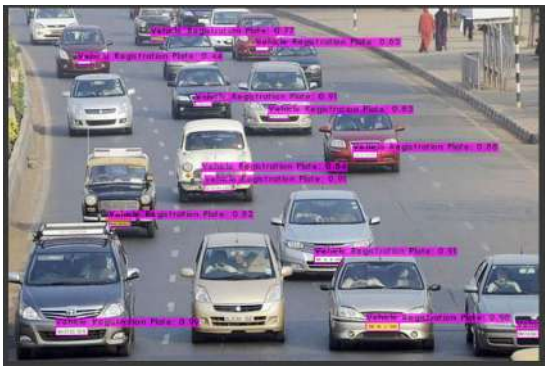


Fig. 5 Result showing the labelling of Number Plates for vehicles after training and detection using YOLOv4 technique over Open Image Dataset.

After overall analysis and comparing the accuracies of YOLOv3 and YOLOv4 models, we have used the YOLOv4 model in our License Plate Detection because of it's higher accuracy and the subsequent result of the detection model is passed to recognition model which recognizes the alphanumeric characters. The output of the same is shown in Fig. 6.

Fig. 6 Result showing accurate detection and recognition of the number plate from a video.

VII. CONCLUSION

In this paper, the study of techniques for number plate detection and recognition are presented along with implementation. Vehicle plate image is obtained by digital cameras, recognition method is applied to it and the image is processed to get the vehicle number plate information. The implementation is done in accordance with the maximum referred research papers. Image is captured from the video by using means of defining a code that extracts keyframes. After keyframes have been extracted, they are passed to the Yolo v4 model which detects the number plate from the extracted frames.

The dataset used is Vehicle Registration Plate Dataset by Google's Open Image Dataset. 850 images are used for training YOLO models and 150 images are used for testing. An accuracy of 92.5% is achieved from Yolo v3 whereas 97.73% is achieved from the YOLO v4 model in number plate detection as shown in Table 2. These images are segmented and then passed on to OCR where character recognition takes place through Tesseract. Once the characters are recognized they are displayed as output.

VIII. REFERENCES

[1] Q. Liu, B. Liu, Y. Wu, W. Li and N. Yu, "Real-Time Online Multi-Object Tracking in Compressed Domain," in IEEE Access, vol. 7, pp. 76489-76499, 2019.

[2] Y. Xu, X. Zhou, S. Chen and F. Li, "Deep learning for multiple object tracking: a survey," in IET Computer Vision, vol. 13, no. 4, pp. 355-368, 2019.

[3] Chirag Patel, Dipti Shah, Atul Patel, “Automatic Number Plate Recognition System(ANPR): A Survey”, International Journal of Computer Application, vol. 69, no. 9, pp. 0975 – 8887, May 2013.

[4] Anisha goyal, Rekha Bhatia, “Automated Car Number Plate Detection System to detect far number plates”, Journal of Computer Engineering, vol. 18, no. 4, pp. 34 - 40, July 2016.

[5] R. Chauhan, K. K. Ghanshala and R. C Joshi, "Convolutional Neural Network (CNN) for Image Detection and Recognition", First International Conference on Secure Cyber Computing and Communication (ICSCCC), pp. 278-282, December 2018.

[6] Dhiraj Y. Gaikwad, Pramod B. Borole, “A Review Paper on Automatic Number Plate Recognition System” International Journal of Advanced Trends in Computer Science and Engineering, vol 1, no. 1, pp. 88-92, April 2014.

[7] Er. Kavneet Kaur, Vijay Kumar Banga, “Number Plate Recognition using OCR Technique”, International Journal of Research in Engineering and Technology, vol. 2, no.9, pp. 286-290, September 2013.

[8] Aniruddh Puranic, Deepak K. T., Umadevi V., “Vehicle Number Plate Recognition System: A Literature Review and Implementation using Template Matching”, International Journal of Computer Applications, vol.134, no. 1, pp. 12-16, January 2016.

[9] Siddharth U. Mishra, Amit A. Badgujar, Sushant N. Asija, Bhavana Julme, “A Review Paper on Automatic License Plate Recognition System (ALPR) using Enhanced Image Processing Techniques for Criminal Surveillance”, International Journal of Engineering Research & Technology (IJERT) vol. 3, no. 2, pp. 1536-1541, February 2014

[10] Zuhaib Akhtar, Rashid Ali, “Automatic Number Plate Recognition Using Random Forest Classifier", SN Computer Science, Springer Nature Singapore Pte Ltd, pp. 1-120, April 2020.

[11] Sik-Ho Tsang, “Review: YOLOv3 — You Only Look Once (Object Detection), Improved YOLOv2: Comparable Performance with RetinaNet.,” Available [Online] Accessed on 30 April 2021. https://towardsdatascience.com/review-yolov3-you-only-look-once-object-detection-eab75d7a1ba6

[12] Manivannan Murugavel, “YOLO V4 - Speed and accuracy are both improved,” May 4, 2020. Available [Online], Accessed on 20 April 2021. https://manivannan-ai.medium.com/yolo-v4-750cd627064f

[13] Alexey Bochkovskiy, Chien-Yao Wang, Hong-Yuan Mark Liao, “YOLOv4: Optimal Speed and Accuracy of Object Detection”, Institute of Information Science Academia Sinica, Taiwan, 23 April 2020.

[14] Mrs. J. V. Bagade, MSukanya Kamble, Kushal Pardeshi, Bhushan Punjabi, Rajpratap Singh, “Automatic Number Plate Recognition System: Machine Learning Approach”, IOSR Journal of Computer Engineering (IOSR-JCE), ISBN: 2278-8727, pp. 34-39.

IDENTIFICATION OF MENTAL HEALTH RELATED ISSUES FROM SOCIAL MEDIA USING NATURAL LANGUAGE PROCESSING

Midhun VM[#1], Sharanya Menon[#2], Akash Patil[#3], Prof. Dhiraj Amin[#4]

Department of Information Technology,

Pillai College of Engineering, New Panvel-410206

Maharashtra, India.

[1,2,3]Student, IT Engineering, Pillai College of Engineering, Maharashtra, India

[4]Faculty, IT Engineering, Pillai College of Engineering, Maharashtra, India

---

Abstract- **Mental illness is one of the most pressing public health issues of our time. While counseling and psychotherapy can be effective treatments, our knowledge about how to conduct successful counseling conversations has been limited due to lack of large-scale data with labeled outcomes of the conversations.Social Media posts contain various types of topics in our daily life, which include health-related topics. Analysis of health-related social media posts would help us understand health conditions and mental health issues encountered by people. The approach is to extract casualties from social media platforms like twitter using natural language processing (NLP) techniques.The basic idea is to create a system that can analyse the data extracted from social media platforms to interpret a person's mental health. This data can be extracted from as many social media platforms as possible to increase the chances of getting an accurate result. Natural Language Processing will be used to analyse the words used in each post. Based on the result the person can be advised as to what must be their next approach to better mental health condition. The system will try to find certain keywords like 'stress', 'melancholy', 'insomnia', 'sad', etc., in the social media posts and try to identify the mental health condition of the user and the help he or she needs.**

Keywords- **Mental Illness, Natural Language Processing, Mental Health, Social Media, Text Mining, Artificial Intelligence**

I. INTRODUCTION

Mental health includes our emotional, psychological, and social well-being. It affects how we think, feel, and act. It also helps determine how we handle stress, relate to others, and make choices. Mental health is important at every stage of life, from childhood and adolescence through adulthood.

Over the course of your life, if you experience mental health problems, your thinking, mood, and behavior could be affected. Many factors contribute to mental health problems, including:

- Biological factors, such as genes or brain chemistry
- Life experiences, such as trauma or abuse
- Family history of mental health problems

Everyone has some risk of developing a mental health disorder, no matter their age, sex, income, or ethnicity. Social and financial circumstances, biological factors, and lifestyle choices can all shape a person’s mental health. A large proportion of people with a mental health disorder have more than one condition at a time. It is important to note that good mental health depends on a delicate balance of factors and that several elements of life and the world at large can work together to contribute to disorders.

According to the World Health Organization (WHO): “Mental health is a state of well-being in which an individual realizes his or her own abilities, can cope with the normal stresses of life, can work productively, and is able to make a contribution to his or her community.” People tend to forget the fact that mental health is equally important to physical health. Studies show that mental health can directly affect the physical health of a person as well. The mental health of a person can be suggestive of their immunitive. High stress can lead to poor immunity.

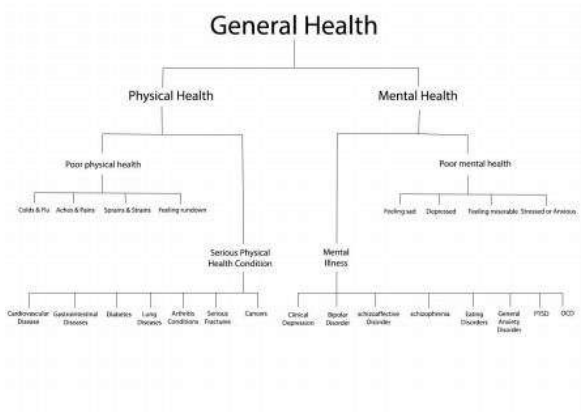


Fig 1. General Health spectrum of a person

The figure 1.1 shows the General Health spectrum of a human being and as we can see half of it focuses on mental health. This is plenty of evidence proving the fact that mental health is equally important.

II. RELATED WORK

Extract health-related causality from Twitter messages using Natural Language Processing: Son Doan, Elly W.Yang, Sameer S.Tilak, Peter W. Li, Daniel S. Zisook and Manabu Torii (April 2019)

In this report an NLP approach was made to extract causality from Twitter messages. They collected the results worth four months of twitter posts related to health and used them to analyse the topics of interests in health. Specifically, the number of matched sentences was 501 out of 29,705 for stress(1.6%), 72 out of 3827 for insomnia(1.8%), 94 out of 11,252 for headache(0.8%). The final casualties extracted were 41 for insomnia, 98 for stress, and 42 for headache. The study had a few limitations as well. The number of rules and patterns are small and they may miss some cause-effect relations in sentences. This study only considered simple cases explicitly reported in single sentences.

A Comprehensive Study on Social Network Mental Disorders Detection via Online Social Media Mining: Hong-Han Shuai , Chih-Ya Shen, De-Nian Yang , Senior Member, IEEE, Yi-Feng Carol Lan, Wang-Chien Lee, Member, IEEE, Philip S. Yu, Fellow, IEEE, and Ming-Syan Chen (July 2018)

In this report,online mining social behavior provides an opportunity to actively identify SNMDs at an early stage. Instead, a machine learning framework, namely, Social Network Mental Disorder Detection (SNMDD), that exploits features extracted from social network data to accurately identify potential cases of SNMDs. It also exploits multi-source learning in SNMDD and proposes a new SNMD-based Tensor Model (STM) to improve the accuracy. Our framework is evaluated via a user study with 3,126 online social network users. We conduct a feature analysis, and also apply SNMDD on large-scale datasets and analyze the characteristics of the three SNMD types. The results manifest that SNMDD is promising for identifying online social network users with potential SNMDs.

Natural language processing to extract symptoms of severe mental illness from clinical text: the Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) project: Richard G Jackson, Rashmi Patel, Nishamali Jayatilleke, Anna Kolliakou, Michael Ball, Genevieve Gorrell, Angus Roberts, Richard J Dobson, Robert Stewart (Jan 2017)

In this report, NLP is used to create a language model to capture key symptoms of Severe Mental Illness(SMI) from clinical texts. This is done by development and validation of information extraction applications to make sure of SMI symptoms. The distribution of derived symptoms was described in 23 128 discharge summaries from 7962 patients who had received an SMI diagnosis, and 13 496 discharge summaries from 7575 patients who had received a non-SMI diagnosis. Data extracted from 46 symptoms with median score of 0.88, four poorly performed models were excluded.

Detecting depression and mental illness on social media: an integrative review: Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar and Johannes C Eichstaedt (2017)

This paper reviews the recent studies on predicting mental illness through social media. With an improved diagnosis rate of mental illness, some cases might remain undetected. These cases can be found on social media such as Twitter and Facebook. These cases are identified based on their activities on various social platforms, online forum membership, screening surveys and patterns in their language. Large-scale monitoring of social media through automated detection methods also helps to identify depression and other mental illnesses.

Natural language processing in mental health applications using non-clinical texts: Rafael A Calvo, David Nicolas Milne, M. Sazzad Hussain, Helen Christensen (Nov 2016)

This paper gives a taxonomy of data sources and techniques used for mental health support and intervention.

Using Facebook, Twitter and other social media to create an online pathway to direct people to health information. It also aims in providing mental health assistance and to generate personalised intervention. Here, the use of social media as a data source which in turn is used to detect emotions and identify people in need of psychological assistance is taken into consideration. Various techniques that are used in labelling and diagnosis is mentioned. Paper also includes wayse aligned to generate and personalise mental health interventions along with the aim to develop a common language that helps to deal with mental health without regional boundaries.

III. PROPOSED SYSTEM

The architecture of the project that we propose is as follows:

Here, the process begins with the collection of raw data from various social networking sites. This data is then pre-processed. This includes cleaning the data that is acquired from the sources and to remove errors from the data. The irrelevant data collected is also removed during this process. It is done to identify and remove missing data and to reduce the original data so as to use the meaningful and reliable data set (Feature Extraction) for better results.

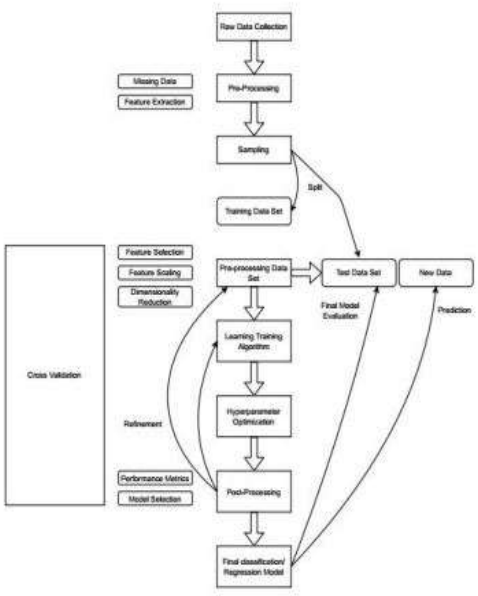


Fig 2. Proposed system architecture

Some common feature extraction techniques are:

- Principal Components Analysis (PCA) :- PCA is one of the most used linear dimensionality reduction techniques. When using PCA, we take as input our original data and try to find a combination of the input features which can best summarize the original data distribution so that to reduce its original dimensions. In PCA, our original data is projected into a set of orthogonal axes and each of the axes gets ranked in order of importance.
- T-distributed Stochastic Neighbour Embedding (t-SNE) :- t-SNE is non-linear dimensionality reduction technique which is typically used to visualize high dimensional datasets. It is extensively applied in image processing, NLP, genomic data and speech processing. It minimizes the divergence between two distributions: a distribution that measures pairwise similarities of the input objects and a distribution that measures pairwise similarities of the corresponding low-dimensional points in the embedding.

This pre-processed data is then subjected to sampling after which it is split into training and testing data sets. Training data set is again pre-processed to perform:

- Feature Selection :- It is a process of reducing the number of input variables when developing a predictive model. It is desirable to reduce the number of input variables to both reduce the computational cost of modeling and, in some cases, to improve the performance of the model.
- Feature Scaling :- It is a method used to normalize the range of independent variables or features of data.
- Dimensionality Reduction :- It reduces the time and storage space required. Removal of multicollinearity improves the interpretation of the parameters of the machine learning model. It becomes easier to visualize the data when reduced to very low dimensions such as 2D or 3D.

This data set is then subjected to the learning algorithm. Hyperparameter optimization is done, that is, hyperparameter optimization or tuning is the problem of choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a parameter whose value is used to control the learning process.

- Training Loss : Training loss or Loss is the error on the training set of data.
- Validation Loss: Validation Loss is the error after running the validation set of data through the trained network. Train/valid is the ratio between the two. Unexpectedly, as the epochs increase both validation and training error drop.

Given below is the graphical representation of training loss and validation loss of the model.

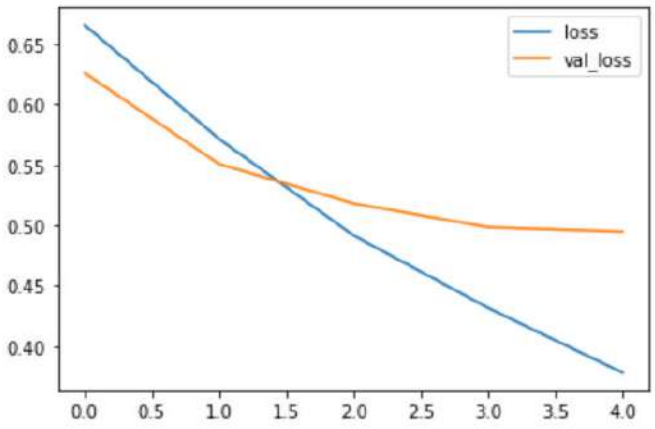


Fig 3. Loss and Validation Loss

Explanation of the terms associated with confusion matrix are as follows −

1. True Positives (TP) − It is the case when both actual class & predicted class of data point is 1.
2. True Negatives (TN) − It is the case when both actual class & predicted class of data point is 0.
3. False Positives (FP) − It is the case when the actual class of data point is 0 & predicted class of data point is 1.
4. False Negatives (FN) − It is the case when the actual class of data point is 1 & predicted class of data point is 0.0

- Accuracy :- Accuracy in classification problems is the number of correct predictions made by the model over all kinds predictions made.

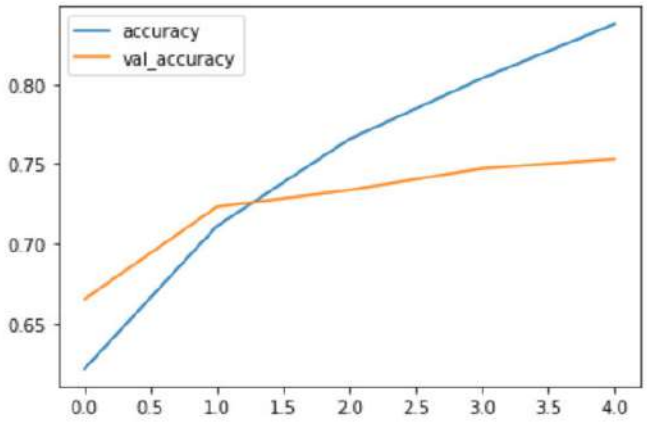


Fig 4. Accuracy and Validation Accuracy

- Precision :- It is defined as the number of correct documents returned by our ML model.

The testing data set is used to test the model and predict the accuracy.

Given below is the snapshot of the application in work. It analyses the text entered by the user and detects whether the user is suffering from anxiety.

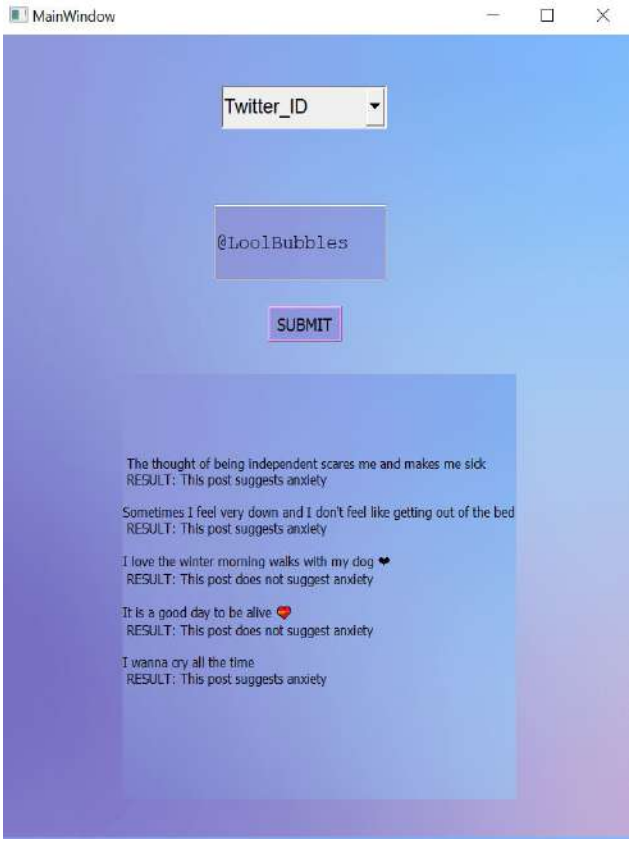


Fig 5. Working of the application

IV. APPLICATIONS

The application is capable of analysing texts and tweets that the user posts and utilising it to predict whether the user has anxiety and needs help alleviate the same. It can be equipped to detect similar mental health problems from the emoticons that the user frequently uses since it has become a huge part of speech while expressing oneself on social media platforms. The application can be improved by inculcating audio analysis of the user to better understand the state of mind of the users from their voice. It can also be equipped to perform identification of complex facial features and expressions using emotion detection.

V. SUMMARY

Here, we have briefly presented an application for Identifying Mental Health related issues from social media using NLP, we have explained the objectives of the proposed system, we have conducted a literature survey of previous works and at the end of chapter 2 we have summarised the literature survey and listed the the advantages and disadvantages of each research paper. We have briefly explained the existing system architecture and the proposed architecture. Further, the report also explains all the tools and technologies implemented in the proposed system such as PyQT and Tensorflow.

VI. ACKNOWLEDGEMENT

We would like to extend our deepest gratitude to our Project Guide, Prof Dhiraj Amin, who guided us and provided us with his valuable knowledge and suggestions on this project and helped us improve our project beyond our limits. Secondly, we would like to thank our Project Coordinator, Prof Krishnendu Nair, who helped us finalize this project within the limited time frame and by constantly supporting us.We would also like to express our heartfelt thanks to our Head of Department, Dr. Satishkumar Verma, for providing us with a platform where we can try to work on developing projects and demonstrate the practical applications of our academic curriculum.We would like to express our gratitude to our Principal, Dr. Sandeep Joshi, who gave us a golden opportunity to do this wonderful project on the topic of 'Identification of mental health related issues from social media using Natural Language Processing', which has also helped us in doing a lot of research and learning their implementation.

REFERENCES

[1]Government of India Mental Health Article
Last Accessed on November 2020

[2] Medical News today article on Risk Factors for Mental Health conditionsLast Accessed on October 2020

[3] Son Doan, Elly W.Yang, Sameer S.Tilak, Peter W. Li, Daniel S. Zisook and Manabu Torii, "*Extract health-related causality from Twitter messages using Natural Language Processing*", BMC Medical Informatics and Decision Making, Article number: 79, April 2019.

[4] Richard G Jackson, Rashmi Patel, Nishamali Jayatilleke, Anna Kolliakou, Michael Ball, Genevieve Gorrell, Angus Roberts, Richard J Dobson, Robert Stewart, "*Natural language processing to extract symptoms of severe mental illness from clinical text: the Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) project*", Jan 2017.

[5] Hong-Han Shuai , Chih-Ya Shen, De-Nian Yang , Senior Member, IEEE, Yi-Feng Carol Lan, Wang-Chien Lee, Member, IEEE, Philip S. Yu, Fellow, IEEE, and Ming-Syan Chen, "*A Comprehensive Study on Social Network Mental Disorders Detection via Online Social Media Mining*", vol 30, no. 7, July 2018.

[6] Rafael A Calvo, David Nicolas Milne, M. Sazzad Hussain, Helen Christensen, "*Natural language processing in mental health applications using non-clinical texts*", received Jan 2016, revised Nov 2016, accepted Nov 2016.

[7] Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar and Johannes C Eichstaedt, "*Detecting depression and mental illness on social media: an integrative review*", 2017.

BIOGRAPHIES

Midhun VM,
Student of Pillai College of Engineering, New Panvel.
Pursuing Bachelor's of Engineering Degree in IT Engineering from University of Mumbai.

Sharanya Menon,
Student of Pillai College of Engineering, New Panvel.
Pursuing Bachelor's of Engineering Degree in IT Engineering from University of Mumbai.

Akash Patil,
Student of Pillai College of Engineering, New Panvel.
Pursuing Bachelor's of Engineering Degree in IT Engineering from University of Mumbai.

Dr Dhiraj Amin,
Professor at Pillai College of Engineering, New Panvel,
Department of IT Engineering.

Advance e-Tutor:'E-programming hut' Based on E-Learning

Priyesh Patil
patilpk43@student.mes.ac.in

Supriya Thale
thalesr15it@student.mes.ac.in

Nikhil Suryavanshi
nikhilps469@student.mes.ac.in

Prof. Krishnendu Nair knair@mes.ac.in

Rohan Vengurlekar
rohansuven16de@student.mes.ac.in

Information Technology Pillai College Of Engineering, Panvel,,Navi Mumbai,India

Abstract: The issue of e-learning as an advance system for training and educating number of people using information and communication technology. It has been received an increasing level of interest in recent few years. This report overcomes design and development issues of an online education for enhancing programming skills of the user.Existing system "E-tutor" has an issue of interaction between student and teacher as in the traditional classroom.It is difficult to solve the doubt for student if they don't have any interaction with the instructor or teacher. The purpose is to design an "advanced layer based e-tutor : "E-programming hut" based on elearning". This system will provide an efficient knowledge of programming and will help user to enhance their programming skills.It will also provide certificate to the student who will complete the course successfully and passes the exam.

Keywords: E-learning,eTutor, E-Programming Hut, knowledge, syllabus outside on going time-honored **programming, certificate**

I. Introduction

E-Learning is learning where we can use electronic technology to access educational lecture room. It can refers to a course, program or degree delivered completely online. E-learning includes the use of a computer or electronic device e.g. a mobile phone in same manner to provide training, educational or learning material.

II. Domain Techniques

The classification of various techniques the domain is given in Figure 1

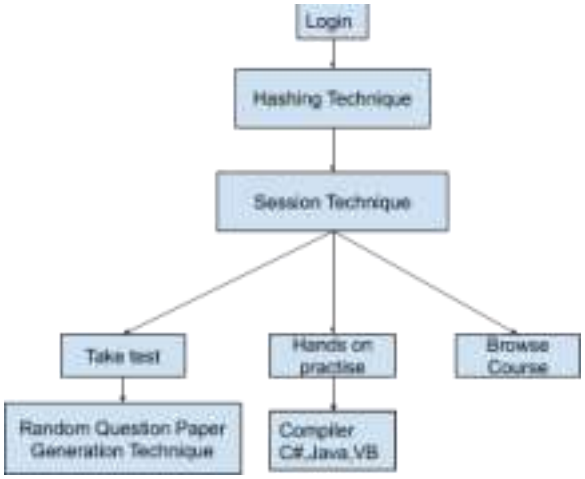


Fig. 1 Classification of domain techniques This system helps user to find information by providing them with personalized suggestions. Based on above problems of researchers, recommendation techniques will have great influence in all aspects of our life.

Existing System:

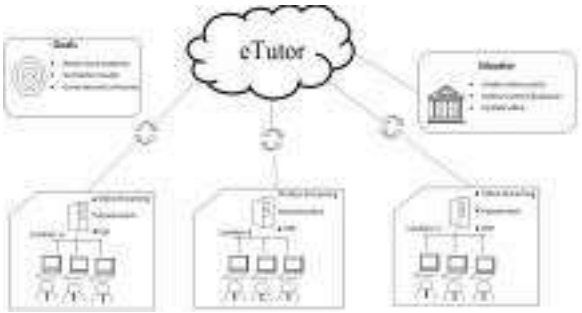


Fig. 2 Existing system used for Content based Systems

Above figure which is an online web-based education system and learns how to teach a course,a concept or remedial materials toa student with specific context in the most efficient way.

Basically for the current student,eTutor learns from its past interaction with students with similar context,the sequence of teaching material that are show to these student,and the response of these student to the teaching material including the final exam scores,how to teach the course in most effective way.This is done by defining a teaching effectiveness metric, referred to as the regret, that is the function of the final exam score and time cost of teaching to the student, and then designing a learning algorithm that learns to optimize this metric. This tradeoff between learning (exploring) and optimizing (exploiting) is captured by the eTutor in the most efficient way, i.e., the average exam score of the students coverage to the average exam score that could be achieved by the best teaching strategy. We illustrate the efficiency of the proposed system in a real-world experiment carried out on students in a DSP class.The following fig 3

Proposed System Architecture

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

In general, hybrid recommenders are systems that combine multiple recommendation techniques together to achieve a synergy between them. Although there exist a number of recommendation approaches that are practical to merge (i.e. Collaborative, Content-based, Demographic and Knowledge-based Recommender), our work will mainly focus on the combination of CF and CBF techniques. The proposed architecture is shown in Figure 3.3

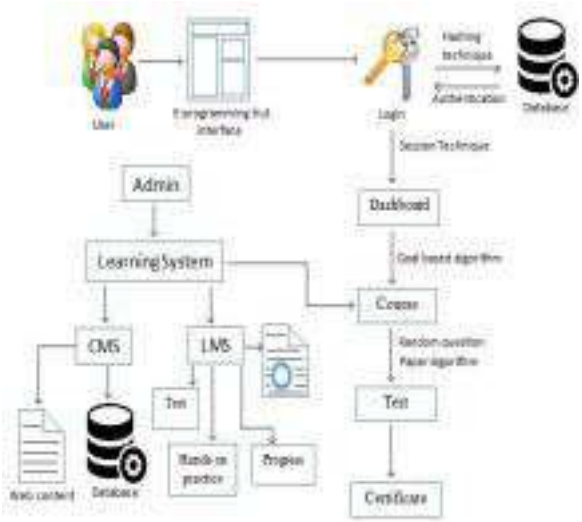


Fig. 3 Proposed system architecture

There are three main entity's in E-Learning system are as follows:

The administrator, instructor, or user requests the login page and enters the username and password. The system verifies the credentials and checks whether the user

is authorized or not. If the user is authorized then he can access the pages, otherwise a notification will be displayed to inform that the access is failed. The web classes contain the methods and functions used to access the database. The web configuration component is used to configure the system functionality and to specify the application settings.

Requirement for implementation Techniques

1. Session Technique:

Session Tracking is a way to maintain state (data) of an user. It is also known as session management in servlet. Http protocol is a stateless so we need to maintain state using session tracking techniques. Each time user requests to the server, server treats the request as the new request. The techniques in this category are adapted to the individual needs, interests and preferences of user or society. They are tools for suggesting items to users in this domain. Various techniques in this category are listed here. These techniques have various advantages and are used extensively in literature.

Basically there are four techniques which can be used to identify a user session.

a. Cookies
b. Hidden Fields
c. Session Tracking API

Cookies, Hidden Fields involves sending a unique identifier with each request and servlets determines the user session based on the identifier. Session API uses the other three techniques internally and provides a

session tracking in much convenient and stable way.

a. Cookie

Cookie is a key value pair of information, sent by the server to the browser and then browser sends back this identifier to the server with every request there on.

There are two types of cookies:

1. Session cookies - are temporary cookies and are deleted as soon as user closes the browser. The next time user visits the same website, server will treat it as a new client as cookies are already deleted.

The following fig 3.4a [8]

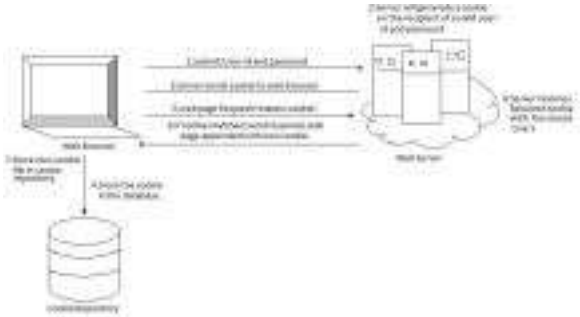


Fig. 3.4(a) Cookie for session tracking

b. Hidden Field

Hidden fields are the input fields which are not displayed on the page but its value is sent to the servlet as other input fields. For example

<input type=”hidden” name=”sessionId” value=”unique value”/>

is a hidden form field which will not displayed to the user but its value will be send to the server and can be retrieved using request.getParameter(“sessionId”) in servlet.

c. Session Tracking API

Servlets provide a convenient and stable session-tracking solution using the HttpSession API. This interface is built on the top of above discussed approaches. Session tracking in servlet is very simple and it involves following stepsGet the associated session object (HttpSession) using request.getSession().To get the specific value out of session object, call getAttribute(String) on the HttpSession object.To store any information in a session call setAttribute(key,object) on a session object.To remove the session data , call removeAttribute(key) to discard a object with a given key.To invalidate the session, call invalidate() on session object. This is used to logout the logged in user.

2. Automated question paper generation Examination is the process which tests users on what knowledge they have gained during the course of time. This exams tells us how much the user have stacked the e-knowledge along with hands on practise and can remember it for a longer period of time. To test the users on the course they have done, a set of multiple choice questions have been added to the database by the instructor. This database is called when the user finishes the course completely, i.e. the user has to complete the course 100% only then, the

user can appear for the exam. Once the user applies for the exam, user will have to face few mcq’s. These mcq’s contain theoretical as well as programming related questions. These mcq’s are called on the users exam page in a random format. This is done using the Random Question Paper Generation technique. This technique randomizes the questions and are put forward on the users exam page. After the submission of the exam, the user will come to know the result if the user is passed or failed. Thus the quality of the exam questions produced by the instructor would determine the quality of the students produced by the institutions. Preparing exam questions is a challenge in traditional system, but it is way easy due to random question paper generation algorithm. This technique does not allow duplication or repetition of the question of paper for the users. So this technique generates unique question paper for all the users. This system consists a highly efficient random algorithm which uses an array to store randomly generated numbers. The questions are then selected against these array elements, hence ensuring unique question papers for all the users. This technique can be seen in the following fig 3.5b [10]

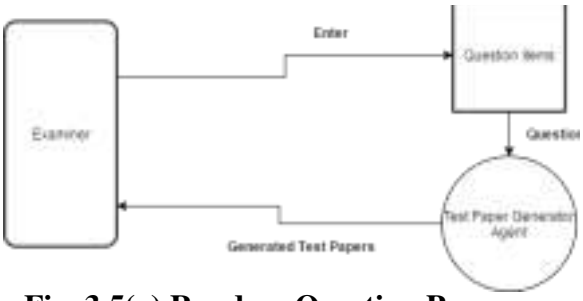


Fig. 3.5(a) Random Question Paper Generation Technique

3. Hashing Technique

Hash algorithms are one way functions. They turn any amount of data into a fixed-length "fingerprint" that cannot be reversed. They also have the property that if the input changes by even a tiny bit, the resulting hash is completely different . This is great for protecting passwords, because we want to store passwords in a form that protects them even if the password file itself is compromised, but at the same time, we need to be able to verify that a user's password is correct.

The general workflow for account registration and authentication in a hash-based account system is as follows: The user creates an account.

Their password is hashed and stored in the database. At no point is the plain-text (unencrypted) password ever written to the hard drive.

When the user attempts to login, the hash of the password they entered is checked against the hash of their real password (retrieved from the database).

If the hashes match, the user is granted access. If not, the user is told they entered invalid login credentials.

- Steps 3 and 4 repeat every time someone tries to login to their account.

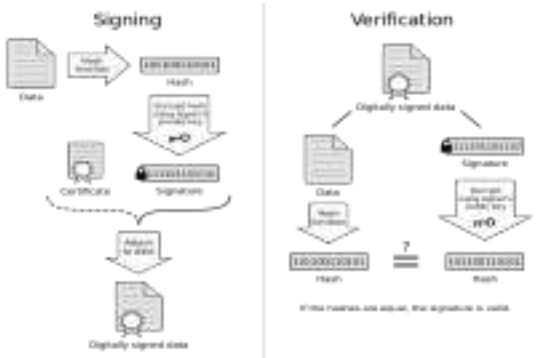


Fig. 3.5(b) Hashing Technique for authentication

4. Goal based algorithm

Goal based algorithm is an algorithm used in e-programming hut that defines a specific goal for the course. The goal for the user is to go through all the chapters, sub-chapters and perform set of exercises for the sub chapters. The user cannot jump on to next chapter or sub-chapter until and unless current chapter and its sub-chapters exercise is completed. Once all the chapters of the course are done, the user can apply for the test. The user can only give the test, if the goal of completing 100% course is achieved. This makes us understand that the user is thoroughly studying the course on our site. The exercises are prepared in such a way that the user has to make changes according to the given conditions and the output of the programs should match the output that is saved in the database. If the output matches, the user can go to the next sub-chapter. Hence the goal is attained, and knowledge is gained by the user.

5. Programming Compiler

When the user is practising java language, the java compiler creates 2 files namely filename.class and filename.java and

stores in the specified paths. When the user is practising c# language, the compiler creates 2 files namely filename.cs and filename.exe. When the user is practising VB language, the compiler creates 2 files namely filename.cs and filename.exe. Then the output of the program is checked with the expected output of the program that is stored in the database. If the output matches, the user can click on finish chapter and can move on to the next chapter or else the user has to get the expected output to get to the next chapter.

The client uses the resources of the server system and in return the server provides services to the client system. These services from the server are possible through iis manager and sql server authentication . IIS manager helps to publish the website on the serve. This server provides localhost ip address to the website to be published. Using this localhost ip address, the client can get access to the website and its server system. The sql server authentication is done by using user id and password of the sql server. The hosting requires data connectivity which is done using sql server authentication by applying its user id and password.

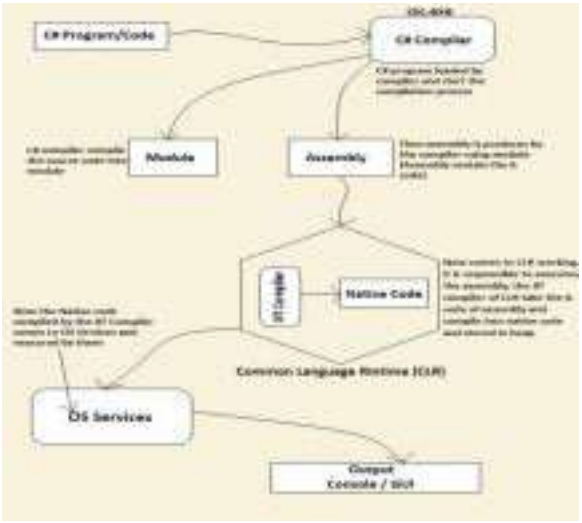


Fig.3.6 C# compiler for compilation of code

Summary

In this report, the study of different domain techniques is presented. The different techniques such as Hashing Technique, Session Technique, Random Question Generator Technique and Goal Based Technique. With the advent of technology, life has become fast-paced. Most of the age-old systems are being upgraded in sync with the latest technological assets. This is to facilitate better learning and provide a hands – on experience. In this project E-learning hut will provide online courses of different languages for user. Languages will be divided into chapters, so that user can easily understand and access the topic. As you have access to the net 24x7, you can train yourself anytime and from anywhere also.

References

[1]Designing E-learning content using AGLOs

Author: Felicia-Mirab ela Costea,Cipria n-Bogdan Chirila,Vladm ir-Ioan Cretu 2019

[2] A Novel Teaching Strategy through Adaptive Learning Activities for Computer Programming

Author:Chris tos Troussas, Akrivi Krouska, and Cleo Sgouropoulou 2020

[3] Implementation of Enhanced Secure Hash Algorithm Towards a Secured Web Portal

Author:Froila n E. De Guzman,Bob by D. Gerardo 2019

[4] Application of Innovative Technologies on the E-learning System

Author:Jianx ia Chen, Qianqian Li,Cleve Yeo Keng Lin,Huapeng Chang,Chunz hi Wang

[5] Online Tools to Support Novice Programming: A Systematic Review

Author: Tze Ying Sim,Sian Lun Lau 2018

[6] Automated analysis of e-learning web applications

Authors:F.Škopljanac-M ačina, B. Blašković i I. Zakarija 2019

Dynamic Traffic Monitoring System

Karan Bhoir ,Vijay Sagar Sekar, Rajkumar Vishwakarma and Prof Mimi Cherian

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

Abstract— Road traffic and traffic congestion are the major problems worldwide, the conventional traffic patterns are nonlinear and complex and time dependent rather than traffic dependent. To avoid such problems traffic monitoring becomes important. The purpose of this project is to develop a system which monitors the traffic using image processing techniques by identifying the numbers of vehicles in a particular lane and allocate time dynamically to that lane to pass and thus reduces the waiting time of the vehicle and it also monitors whether someone is violating the traffic rules. If someone crosses the road when the light is red, the camera clicks the image of the vehicle and extracts the vehicle number and sends an e-chalan to the registered mobile number of the vehicle..

Keywords—Traffic control, Traffic monitoring, Image processing, Camera, Traffic violation , e-challan

1. Introduction

Nowadays vehicular traffic surveillance is an important civilian application to improve road control, intelligent road accident detection and urban congestion. Traffic congestion is becoming a problem in every big city. Traffic congestion wastes money, resources and time of the government. And has a huge impact on the environment.The aim of this project is to develop a system which will reduce the traffic congestion significantly by using different techniques. The Dynamic traffic monitoring system is a real time system that captures the images of the road continuously from the signal, these images are then processed to find the no of vehicles on the street. Depending upon the number of vehicles the time is calculated which is sufficient to pass the traffic in that lane. This system will be updated daily by the historic data it has collected and will improve. The system will also predict the traffic and will adjust the timings according to it.This System is further integrated with e-challan system of RTO. If any vehicle passes the signal when it is red the photo of the number plate is captured and the challan is directly sent to the owner of the vehicle.

2. Literature Survey

A. Design of Dynamic Traffic Signal Control System: The proposed DynamicTraffic Light Control (DTLC) operations have Infrared Sensors mounted on the road to detect frequency of the vehicles. The presence or absence of a vehicle is sensed by the sensor assembly mounted on each road,which acts as an input to the DTLC unit. This input signal indicates the density of vehicles on each road. In this system the basic operations are implemented using Microcontroller 89c51 AT. The output is given in the form of three lights: red green and yellow . Also it includes a feature for emergency cases in that situation: signal turns red for all the roads except on the lane in which the emergency vehicle is present [1].

B. Smart control of traffic lights systems using image processing: This system the work is divided into 3 parts. The first part is to acquire the image from a fixed camera. The second part is to process the captured image using image processing techniques. The third part is controlling the traffic lights using two Arduino UNO boards . The image is first captured and then transferred using usb cable; the further processing is done by using MATLAB . the captured image is converted into grayscale image by eliminating the hue and saturation information while retaining the luminance, using weighted method for further processing and the into binary images that contains only two colors and then the traffic density is calculated using some mathematical operation and according to the density the time is allocated for each lane output is given through three colours of led red ,green and yellow connected to the Arduino UNO boards connected through an computer [2].

C. Smart control of traffic lights systems using image processing: This system proposed a tracking algorithm based on mean shift and a projective Kalman filter and pixel based . The algorithm achieves robust tracking due to the integration of the projection equation of the vehicle onto the image plane of the CCD camera. In particular, the observation function of the projective Kalman filter models the trajectory of vehicles with respect to their

ground distance to the camera. The results showed that both the standard and the projective Kalman filter algorithms achieve robust tracking at a rate of 30fps, even though the projective Kalman filter performs better on long distance vehicles Image processing is developed for use as traffic control sensor to obtain multi lane traffic volume, queue length and downstream congestion that may obstruct traffic flow exiting [3].

D. Traffic Congestion Investigating System by Image Processing from CCTV camera: The system uses an image processing technique to analyze for a traffic condition. It detects how many objects or cars are on the road. And then, the system connects a traffic condition result with a database for transportation planning. Moreover, it can be used with other systems such as a traffic light control system on the intersection. There are 3 kinds of results to notify for a traffic condition as flow, heavy and jammed [4].

3. Proposed Work

Creating a dynamic monitoring system through which we can monitor the traffic. In the system we calculate the number of vehicles on one side of the road , we will repeat the same process for the other side of the road. Then we will dynamically allocate time according to the traffic on the road. We will capture the number plate of the vehicles which violates the traffic signals and rules and we issue an e-challan for that vehicle.

3.1 System Architecture

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

A. Image Capture: In this system, the image is captured from the camera and it is then processed forward to the image processing.

***B. Image* Processing:**. The image captured is then passed to the YOLO V3 trained model. Which identifies the different vehicles like cars, bikes, trucks and ambulance. This data is stored for the further calculations. This model also checks whether any vehicle is violating any rules. Different techniques and filters are used to clean the image for the maximum accuracy. The vehicles are marked by rectangles around them with the label

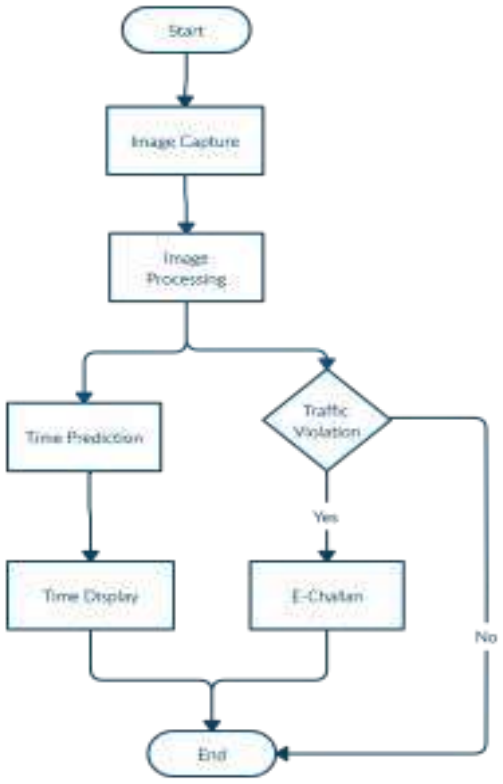


Fig. 1 Proposed system architecture

C. Time Prediction: This Application we need to provide the image of the lane with the traffic at that moment or either we need a camera that clicks the image when it is triggered. The image is the processed using using machine learning algorithms like yolo v3 and yolo v4 .It identifies the different vehicles and present the image it draws a box around the detected abject and the then labels it.The image is stored in the database with the location and the time when the image was captured. Then the program counts the different type of vehicles in the image it is stored in the tuple this tuple is passed as an argument to the model trained using regression algorithm with the dataset of the vehicles at a particular time using this model the time is predicted and the time is displayed on the LED display.

D. Time display: The Time prediction will then be used , so that dynamic time will be displayed on the screen which the vehicle owner can see and drive according to it.

E. Traffic Violation: In this process the same image captured first is passed as an argument.In this part line is

drawn on the image if the vehicle is detected ahead of the line then the vehicle detected is cropped then the cropped image is passed to the number plate detection function In this program the number plate is detected and then the vehicle number is extracted using Tesseract OCR algorithm and the Fine notification is send to the email Id registered with that number also for motorbike detected it checks for number of persons on that motorbike and if the driver is wearing helmet is detected the rules violated the also the fine notification is sent to the registered email with the vehicle number.

F. E-challan: **The** vehicle which violates the traffic rules and signals will be fined according to rules of traffic violation and an e-challan will be sent to the person who violates traffic rules.

3.2. Algorithm:

YOLO : Object detection is one of the classical problems in computer vision where you work to recognize what and where — specifically what objects are inside a given image and also where they are in the image. The problem of object detection is more complex than classification, which also can recognize objects but doesn't indicate where the object is located in the image. In addition, classification doesn't work on images containing more than one object.

YOLO uses a totally different approach. YOLO is a clever convolutional neural network (CNN) for doing object detection in real-time. The algorithm applies a single neural network to the full image, and then divides the image into regions and predicts bounding boxes and probabilities for each region. These bounding boxes are weighted by the predicted probabilities.

YOLO is popular because it achieves high accuracy while also being able to run in real-time. The algorithm "only looks once" at the image in the sense that it requires only one forward propagation pass through the neural network to make predictions. After non-max suppression (which makes sure the object detection algorithm only detects each object once), it then outputs recognized objects together with the bounding boxes.

With YOLO, a single CNN simultaneously predicts multiple bounding boxes and class probabilities for those boxes. YOLO trains on full images and directly optimizes detection performance.

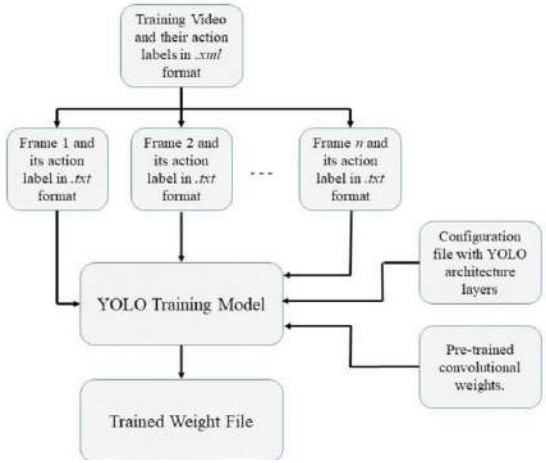


Fig 2 training of object detection model

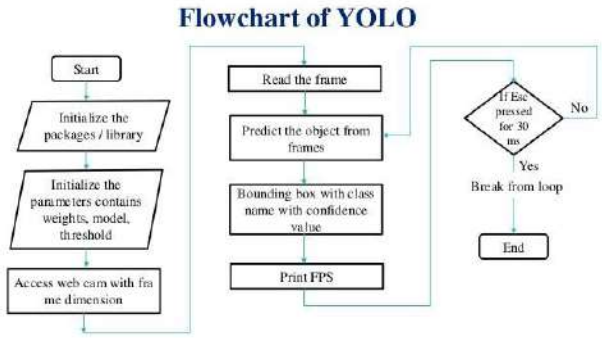


Fig 3.working of YOLO

5. Screenshots of working project

. ACKNOWLEDGMENT

It is our privilege to express our sincerest regards to our BE Project mentor Prof. Mimi Cherian 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. Conclusion

In order to reduce the number of traffic which result in more time for traveling which eventually results in loss of time and money. This gives dynamic monitoring of signal and hence reduction in pollution and saving for the fuel are the most important advantages of this system. Most hurdles of traffic issues will be solved with this system which is cost effective and simple and it makes our life better, safe and time saving. The outcome of this project can be further applied in different applications to give an IOT based solution under different circumstances.

References

[1]Density Based Smart Traffic Control System Using Canny Edge Detection Algorithm for Congregating Traffic Information
https://sci-hub.tw/https://ieeexplore.ieee.org/abstract/document/8275131

[2]An Efficient Algorithm for Detecting Traffic Congestion and a Framework for Smart Traffic Control System
https://sci-hub.tw/https://ieeexplore.ieee.org/abstract/document/7423566

[3]An Efficient Algorithm on Vehicle License Plate Location
https://sci-hub.tw/https://ieeexplore.ieee.org/abstract/document/4636370

[4]Intelligent Traffic Control System (ITCS)
https://ieeexplore.ieee.org/abstact/document/8687368

[5]Smart traffic light control system
https://ieeexplore.ieee.org/abstact/document/7470780

[6] Dynamic traffic monitoring system(IJRIT)
Paper id :- IJSARTV7I443357

MUSIC RECOMMENDATION SYSTEM USING MACHINE LEARNING

Varsha Verma, Ninad Marathe, Parth Sanghavi, and Dr. Prashant Nitnaware Department of Information Technology, PCE, Navi Mumbai, India - 410206

***Abstract*—**

In our project, we will be using a sample data set of songs to find correlations between users and songs so that a new song will be recommended to them based on their previous history. We will implement this project using libraries like NumPy, Pandas.We will also be using Cosine similarity along with CountVectorizer. Along with this,a front end with flask that will show us the recommended songs when a specific song is processed.

***Keywords*—numpy,pandas,cosine similarity,count vectorizer**

I. Introduction

With the explosion of networks in the past decades, the internet has become the major source of retrieving multimedia information such as video, books, and music, etc. People have considered that music is an important aspect of their lives and they listen to music, an activity they engage infrequently. People sometimes feel it is difficult to choose from millions of songs. With commercial music streaming services which can be accessed from mobile devices, the availability of digital music currently is abundant compared to the previous era. Music service providers need an efficient way to manage songs and help their customers to discover music by giving quality recommendations.

A music recommender system is a system that learns from the user's past listening history and recommends songs which they would probably like to hear in the future. By using a music recommender system, the music provider can predict and then offer the appropriate songs to their users based on the characteristics of the music that has been heard previously. Sorting out all this digital music is very time-consuming and causes information fatigue. Therefore, it is very useful to develop a music recommender system that can search in the music libraries automatically and suggest suitable songs to users. Thus, there is a strong need for a good recommendation system.

Recommendation Systems are everywhere and pretty standard all over the web. Currently, there are many music streaming services, like Pandora, Spotify, etc., which are working on building high-precision commercial music recommendation systems. Amazon, Netflix, and many such companies are using Recommendation Systems. Music recommendation is a very difficult problem as we have to structure music in a way that we recommend the favorite songs to users which is never a definite prediction. In this project, we have designed, implemented, and analyzed a song recommendation system. The one we are going to build is pretty common to what Spotify or Youtube Music uses but much more straightforward. Currently, most of the streaming music systems recommend songs based on Collaborative Filtering and Content-Based filtering techniques.

While collaborative filtering (CF) has been the most common choice in those early days of RS research, approaches based on content-based filtering (CBF) have gained popularity in recent years. In short, collaborative filtering approaches exploit interactions between users and items, e.g., clicks or ratings, which are represented in a user-item (rating) matrix *R*.

Collaborative filtering System: Collaborative does not need the features of the items to be given. Every user and item is described by a feature vector or embedding. It creates embedding for both users and items on its own. It embeds both users and items in the same embedding space. It notes which items a particular user likes and also the items that the users with behavior and likings like him/her likes to recommend items to that user. It collects user feedback on different items and uses them for recommendations. Collaborative filtering is further divided into three subcategories: memory-based, model-based, and hybrid collaborative filtering.

Content-based recommendation system: CBRS recommends items based on their features and the similarity between elements of other items. Assuming a user has already seen a movie from the genre of Comedy, CBRS will recommend movies that also belong to the Comedy genre. A content-based recommender works with data that the user provides, either explicitly (rating) or

implicitly (clicking on a link). Based on that data, a user profile is generated, which is then used to make suggestions to the user. As the user provides more inputs or takes actions on the recommendations, the engine becomes more and more accurate.

Python is increasingly being used as a scientific language. Matrix and vector manipulations are extremely important for scientific computations. Both NumPy and Pandas have emerged to be essential libraries for any scientific computation, including machine learning, in python due to their intuitive syntax and high-performance matrix computation capabilities.

NumPy stands for 'Numerical Python' or 'Numeric Python'. It is an open-source module of Python which provides fast mathematical computation on arrays and matrices. Since arrays and matrices are an essential part of the Machine Learning ecosystem, NumPy along with Machine Learning modules like Scikit-learn, Pandas, Matplotlib, TensorFlow, etc. complete the Python Machine Learning Ecosystem.

Flask uses the metropolis example engine to dynamically build hypertext markup language pages mistreating acquainted Python ideas like variables, loops, lists, and so on. We've used these templates as a part of this project. Flask may be a small internet framework written in Python. it's classified as a microframework; as a result of it doesn't need explicit tools or libraries. Templates are files that contain static knowledge still as placeholders for dynamic knowledge. An example is rendered with specific knowledge to provide a final document. Flask uses the metropolis example library to render templates that we've employed in the project. The tactic attribute of type|The shape} component tells the net browser a way to send form knowledge to a server. Specifying a price of POST suggests that the browser can send the info to the net server to be processed. This is often necessary once adding knowledge to info, or once submitting sensitive data. within the project POST technique is employed to require needed song name input from the user, then it's processed into the particular cubic centimeter program for recommending the song. In Python, a lambda perform may be a single-line perform declared with no name, which might have any variety of arguments. However it will solely have one expression. Mistreatment associate degree idles perform and together with the specified variables and lambda perform to fetch the prediction of song recommendation from the cubic centimeter engine for Song Recommendation.

II. Literature Survey

Personalized Recommender Systems

Personalization issues adapting to the individual desires, interests, and preferences of every user. They're tools for suggesting things to users.

Content-based Recommender Systems

Pasquale Lops, Marco American state Gemmis, and Giovanni Semeraro, 2010 [1] in their paper Content-based Recommender Systems: State of the Art and Trends discusses the most problems associated with the illustration of things, ranging from easy techniques for representing structured information to a lot of complicated techniques returning from {the information|the knowledge|the information} Retrieval analysis space for unstructured data.

This work is split into three components. The primary half presents the essential ideas of content-based recommender systems, a high-level design, and their main blessings and disadvantages. The second half a review of the state of the art of systems adopted in many application domains by describing each classical and advanced technique for representing things and user profiles. The foremost wide adopted techniques for learning user profiles also are conferred. The last half discusses trends and future analysis which could lead towards ensuing generation of systems, by describing the role of User Generated Content as how for taking under consideration evolving vocabularies, and also the challenge of feeding users with lucky recommendations, that's to mention amazingly fascinating things that they could not have otherwise discovered.

Hybrid Recommender Systems

Robin Burke, [2] in his survey Hybrid Recommender Systems: Survey and Experiments, explains numerous recommendation techniques. These techniques show the complementary benefits and downsides. It compares the assorted techniques and shows that techniques area unit

higher supported the analysis metrics. This reality has provided an incentive for analysis in hybrid recommender systems that mix techniques for improved performance.
It proposes numerous hybrid approaches which may be accustomed recommendation systems supported the appliance for higher accuracy and results.

Recommendation System Using Association Rules Mining

Luo Zhenghua, 2012 [3] in the realization of individualized recommendation system on book sale applies the association rules in data processing to e-commerce business systems of book sales, styles AN individualized recommendation system of book sales, and introduces the flow of the advice system and therefore the specific realization procedures of information input, knowledge preprocessing, association rules existence and individualized recommendation. Results show that the net website supported this has shown nice performance.

Hybrid Approach for Collaborative Filtering

Gilbert Badaro, Hazem Hajj, Wassim El-Hajj, and Lama Nachman, 2013 [4] in hybrid approach for cooperative filtering for recommender systems talks a couple of new hybrid approach for determining the matter of finding the ratings of unrated things in the user-item ranking matrix by a weighted combination of user primarily based} and item-based cooperative filtering. The projected technique provides enhancements in addressing 2 major challenges of recommender systems: accuracy of recommender systems and scantness of information. The analysis of the system shows the superiority of the answer compared to complete user-based cooperative filtering or item-based cooperative filtering.

The literature survey shows that a hybrid model is projected which mixes user-based cooperative filtering and item-based cooperative filtering by adding the anticipated ratings from every technique and multiplying them with a weight that comes with the accuracy of every technique alone. The approach advantages from the correlation between not solely users alone or things alone however from each at the same time. The analysis was conducted on movielens dataset. the selection of weights was thought of by victimization and adjusting mean absolute error. therefore the survey shows that the hybrid approach improves the information scantness drawback and therefore the accuracy of the system effectively and with efficiency.

Content and collaborative based filtering and association rule mining

Anand Shanker Tewari, Abhay Kumar, and Asim Gopal bartender, [5] proposes a replacement approach to book recommendation system by combining options of content primarily based filtering, cooperative filtering, and association rule mining. The literature survey shows that numerous parameters like content and quality of the book by doing cooperative filtering of ratings by alternative consumers. the aim of this technique is to advocate books to the client that suits their interest. this technique works offline and stores recommendations within the buyer's internet profile. It finds out the class of the book that the client has bought earlier, like a novel, science, engineering, etc. from the consumer's internet profile. It finds out the subcategory of the book.

It performs content-primarily based filtering in class /subcategory, to search out the books that are unit abundant just like the books that the client has bought earlier from the consumer's past history record. On the results of the on top of the step, item primarily based cooperative filtering is performed. This step truly evaluates the standard of the recommending books supported by the rating given to those books by the opposite consumers. From the book dealing info, realize all transactions whose class and subclass are the same as found in step1 and step2.

Non-Personalized Recommender Systems

Non-personalized recommender systems are the only form of recommender systems. They are doing not take into consideration the non-public preferences of the users. The recommendations made by these systems are identical for every client.

Non-Personalized and User-based Collaborative Filtering

Anil Poriya, Neev Patel, Tanvi Bhagat, and Rekha Sharma, Ph. D, [6], in their paper Non-Personalized Recommender Systems and User-based cooperative Recommender Systems describes however websites these days extremely rely on recommender systems. It provides

United States insight into 2 common techniques: non customized recommendation and cooperative filtering. Non Customized recommendations use 2 sorts of algorithms: collective opinion recommender and Basic product association recommender.

The literature review describes, collective opinion recommender that essentially recommends restaurants supported the typical score given to that by different customers. The typical is calculated victimization spherical mean ratings. But these averages lack context throughout recommendations. Thus basic product association recommender is employed. It provides helpful non-personalized recommendations in an exceeding context. Recommendations might not be essentially specific to the user however specific to what the user is presently doing (viewing/buying). The recommendations during this system are similar to all or any users and lack personalization and therefore won't attractive to everybody. Thus cooperative filtering is employed. The cooperative recommender systems overcome the dearth of the personalization involved non-personalized recommender systems. Conjointly no item knowledge is required for this approach and its domain freelance. The machine time is low for model primarily based approaches.

Literature Summary

SN	Paper title	Author & Year of Publication	Methodologies Advantages and Disadvantages
1.	Content-based Recommender Systems	Pasquale Lops, Marco de Gemmis and Giovanni Semeraro. 2010	Advantages: Learning of profile is made easy. Quality improves over time. Considers implicit feedback. Disadvantages: Does not completely Overcome the problem of over-specialization and serendipity.
2.	Hybrid Recommender Systems:	Robin Burke 2010	Advantages: The survey shows combine techniques for improved performance. It improves the user preferences for suggesting items to users.
3.	Association rule Mining for recommendation system on the book sale	Luo Zhenghua. 2012	Advantages: The website based on this has shown great performance. Disadvantages: It does not recommend quality content to the users. Does not consider new user cold start problem Not very efficient in terms of performance
4.	Collaborative filtering for recommender systems: User-based and Item-based CF	Gilbert Badaro, Hazem Hajj, Wassim El-Hajj and Lama Nachman. 2013	Advantages: solves the problem of finding the ratings of unrated items in a user-item ranking matrix. It improves the data sparsity problem. Disadvantage: It does not consider the demographic features which would give better results and solve the user cold-start problem.
5.	Content-Based Filtering, Collaborative Filtering, and Association Rule Mining	Anand Shanker Tewari, Abhay Kumar, and Asim Gopal Barman. 2014	Advantages: It considers various parameters like content & quality of the book by doing collaborative filtering of rating of other buyers. It does not have performance problems. It builds the recommendation offline. Disadvantage: It still lacks the new user cold-start problem.
6.	Non-Personalized Recommender Systems and User-based Collaborative Recommender Systems	Anil Poriya, Neev Patel, Tanvi Bhagat, and Rekha Sharma. 2014.	Advantages: The system helps users find items they want to buy from a business. It overcomes the lack of personalization involved with non-personalized recommender systems. It is domain-independent. Disadvantages: The recommendations are not very specific. It still lacks personalization. The computational time is low.

III. Proposed System

This research focuses on determining the most effective DM technique with the highest precision between the different classification techniques to be used. In addition, finding the effect of train/test data ratio on the accuracy of the prediction.

System Architecture

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

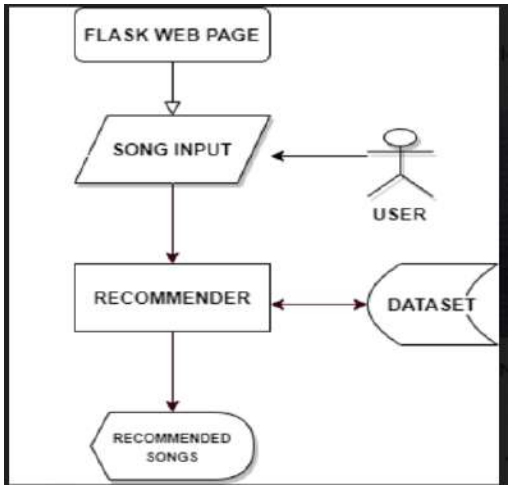


Fig. 1 Proposed System Architecture

A. Data Collection and Understanding Process: The real dataset is used for the research. We have taken music data which contains 2000 records and 15 fields, including categorical and numeric features. Each record in the music data set represents single musical 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 is finished, the process of preparing the data is performed. It is important to refine this data so that it can be suitable for the models and generate better results. In this phase we performed tasks like cleaning, filling the missing data, and removing unwanted data. The data of Spotify had various attributes which were not relevant, i.e., was not giving any useful information, like Title, Artist, Top Genre, Energy, BPM, Liveness, etc.; hence these attributes are removed in this phase.

C. Feature Selection: Feature selection is 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 them are important and these are then used for training methods. From there, we get the top factors that affect performance.

D. Test and Train Dataset:

Separating data into test datasets and training datasets is an important part of evaluating data mining models as it minimizes the effects of data inconsistency and better understands the characteristics of the model. The test data set contains all the required data for data prediction, and the training data set contains all irrelevant data. We have split the dataset into variable ratios to study the estimation of Prediction.

This paper targets getting the most important variables that may positively affect the accuracy of the features of music performance prediction models using the various feature selection algorithms.

IV Modeling and Experiments

Before building the model and software infrastructure, the data preprocessing and cleaning step is done, since the function get important features appends all the required rows, there may come NaN values in the dataset which have to be replaced with an empty string.
We can see the most important features selected in Table .

Sr. No	Attributes
1	Title
2	Artist
3	Top Genre

Table 3 Final Attributes used for prediction

The specified features are the appended to form a long string which is later used to find similarity score for each song.

V. Requirement Analysis

A. Software

The operating systems used will be windows 7& above. Programming languages used are Python, HTML5, CSS3, Bootstrap.

B. Hardware

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. The hard disk drive is required to store the program permanently on the storage. The 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.

VI. Implementation and Result Analysis

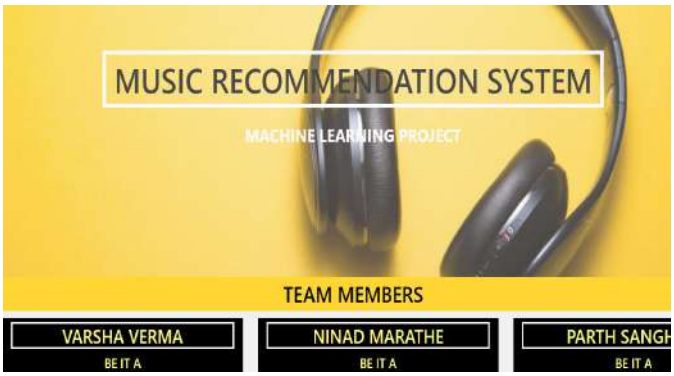
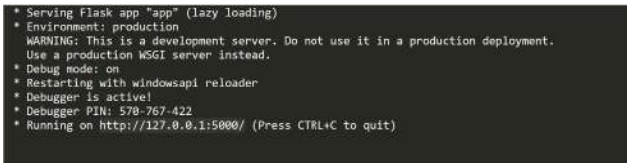
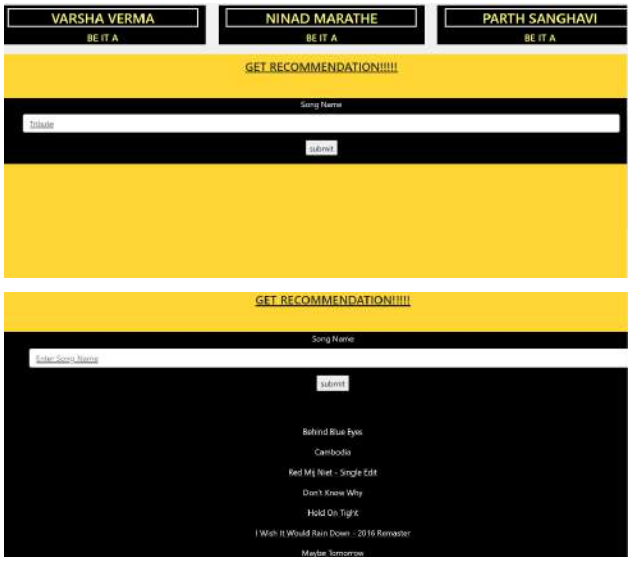


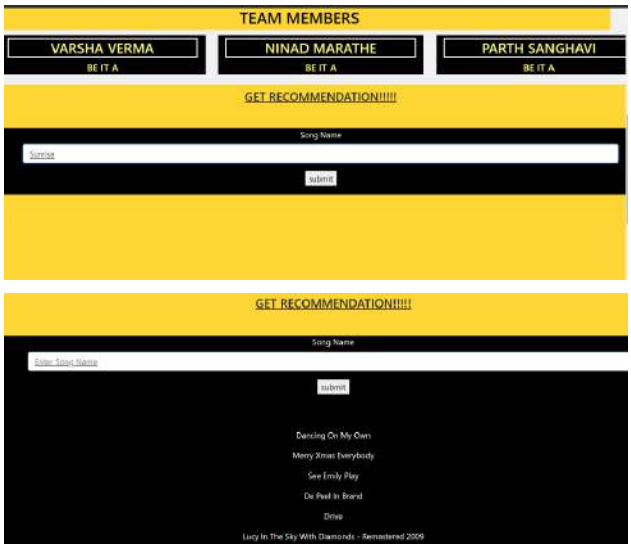
Fig 3 User input page



In the system, First user inputs the song which he/she wants; once the required song is inputted by the user, that ten similar songs are recommended to him. Initially, the process takes into consideration by taking three main features, that is Title, Artist, and Top Genre, which is done by taking Angular distance and Euclidean distance. For this, we have taken the class Count Vectorizer and method cosine similarity. Count vectorizer is stored in an object which is used to count the number of terms that appeared in a particular feature; after that, structured data is used by cosine similarity to find the similarity score. Before the data is processed by the count vectorizer class, since we are using multiple parameters/ features to find the similarity score, a function is created to merge the contents of all the rows of the specified features. In case any NaN values are found, they are replaced with an empty string.



Example figure 1.



Example figure 2

Once we get the cosine similarity between the features, we create a list of enumeration for the similarity score.

After that, the seven most similar songs are predicted by the model which is presented on the frontend.

VII. Conclusion and Future Scope

In the future, we would like to try the following things: 1.Using audio signal (e.g. audio frequency) to recommend songs 2.Trying content-based algorithm 3.Trying Convolutional Neural Network 4.Making the recommender system a real-time system 5, trying clustering techniques to recommend music. Designing a personalized music recommender is complicated, and it is challenging to thoroughly understand the users' needs and meet their requirements. As discussed above, the future research direction will be mainly focused on user-centric music recommender systems. A survey among athletes showed practitioners in sport and exercise environments tend to select music in a rather arbitrary manner without full consideration of its motivational characteristics. Therefore, future music recommenders should be able to lead the users to reasonably choose music. In the end, we are hoping that through this study, we can build the bridge among isolated research in all the other disciplines.

Acknowledgment

It is our privilege to express our sincerest regards to our supervisor Dr. Prashant Nitnaware 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. Sharvari Govilkar, and our Principal, Dr. Sandeep M. Joshi, for encouraging and allowing us to present this work.

References

1. [1] Luo Zhenghua, "Realization of Individualized Recommendation System on Books Sale," IEEE 2012 International Conference on Management of e-Commerce and e-Government. pp.10-13.
2. [2] Tewari, A.S. Kumar, and Barman, A.G, "Book recommendation system based on combining features of content-based filtering, collaborative filtering and association rule mining," International Advance Computing Conference (IACC), IEEE, pp 500 – 503, April 2014.
3. [3] Robin Burke, "Hybrid Recommender Systems: Survey and Experiments", California State University, Department of Information Systems and Decision Sciences, Vol. 12, No. 4, pp. 331-370, March 2012.
4. [4] Anil Poriya, Neev Patel, Tanvi Bhagat, and RekhaSharma, "Non-Personalized Recommender SystemsandUser-basedCollaborative Recommender Systems", International Journal of Applied Information Systems (IJAIS), FCS, Vol. 6, No. 9, March 2014.
5. Fang, J., Grunberg, D., Luit, S., & Wang, Y. (2017, December). Development of a music recommendation system for motivating exercise. In Orange Technologies (ICOT), 2017 International Conference on (pp. 83-86). IEEE.
6. Nakamura, K., Fujisawa, T., & Kyoudou, T. (2017, October). Music recommendation system using lyric network. In Consumer Electronics (GCCE), 2017 IEEE 6th Global Conference on (pp. 1-2). IEEE.
7. Keita Nakamura, Takako Fujisawa. Music recommendation system using lyric network, Journal of 2017 IEEE 6th Global Conference on Consumer Electronics (GCCE), 2017
8. International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-7, May, 2019
9. 4th International Conference on Computer Science and Computational Intelligence 2019 (ICCSCI), 12–13 September 2019
10. Arditi D. Digital Subscriptions: The Unending Consumption of Music in the Digital Era. Popular Music and Society. 2017

Soil Fertigation System for Desired Crop Using IoT and Machine Learning

Divya Dhamankar, Shrutika Ahire, Shahnaz Ussanar, Dhanashree Berde

Prof. Gayatri Hegde

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

***Abstract*—India is a nation of agriculture , its prime importance is to focus on farming and improving the method use for farming. In India agriculture with its allied sectors, is the largest source of livelihoods in india.The agricultural process like irrigation or testing the nutrients content of the soil in terms of fertility taken care of, this is called precision farming. The fertility of the soil is measured by the amount of nutrients present in the soil. There will be two types of nutrients present in the soil are macro and micro nutrients, also water, pH etc.**

In India most of the farmers use the manual technique to use the fertilizers in their agricultural land. Addition of the fertilizers in the right amount is of great importance as excess addition of the fertilizers can harm the plant life and reduce the chance of great yield. Main objective of this project is Soil fertigation system for desired crops using IoT and machine learning. In the proposed system using different sensors measure macronutrients of the soil and transmit the data to the cloud. The user can view the soil fertility at their mobile website. The software system has the intelligence to recommend the fertilisers that are required to be used to suit the needs of the desired crop, thus improving the quality of the soil and in turn, increasing the yield. Overall, the proposed system helps farmers to gather real-time information about various soils, their fertility level, suggest crops and fertilisers at the convenience of the websites.Finally, this project effort will help farmers to make the right decision, gain better yield and economic advantage.

***Keywords*—Smart farming, Irrigation and Fertilization control, Internet of Things, irrigation system.**

1. INTRODUCTION

For farmers soil analysis is the important thing. Soil has a great supply of nutrients. Plants need a controlled environment for healthy growth. Soils are used to continue the growing process so all the nutrients present in the soil get removed whenever the crop is harvested. Low nutrients may contain the deficiency in the crop as well as low production. For healthy crop growth nutrients need to be restored in the soil. When a great amount of the nutrients are present in the soil the plant growth is healthy. So framers need to add the perfect and great amount of nutrients in the soil for healthy crop growth. Organic fertilizers are more important for healthy growth. The nutrients which cannot be replaced by any other elements are called essential nutrients. Which is necessary for crop growth. In nutrients Nitrogen(N), phosphorus (P), potassium(K) are the Main nutrients are in the soil. If we know the right amount of N, P, K needed for the soil. It will help to produce healthy crop growth. With the help of nutrients the plant grows in good conditions The aim of the project is using different wireless sensors to measure the amount of nutrients needed for the soil.

2. LITERATURE SURVEY

[1]. Detection of Nitrogen ,Phosphorus and Potassium nutrients of soil using optical transducer, Marianah Masrie , Mohamad Syamim Aizuddin Rosman, Rosidah Sam and Zuriati Janin ,IEEE 2017.

In this paper [1], optical transducer is used to measure the amount of (N, P, K) content of soil.Optical Transducer makes use of light detection system and provide LCD display control functions ,First uses three LED‟s with different wavelength .LED and photodiode was placed in same length .Light is reflected and detected by the photodiode to detect the length of transducer. Difference in light intensity level will be calculated,absorption rate was calculated and

determined the deficiency of nutrient content in soil in three levels high , low,medium, .Here the displayed values of nutrients are compared with threshold values and then decide the level of nutrients.

[2]. Automated fertigation system for efficient utilization of fertilizer and water, Cyril Joseph, I Thirunavukkarasu, Aadesh Bhaskar, Anish Penujuru,IEEE 2017

This paper[2] gives an implementation of a system consisting of four step soil moisture monitoring and control ,fertilizer mixing and delivery ,Wifi module (ESP826) connection and configurations and user interface. A timer is set for detecting moisture of a soil twice a day.It detects the moisture content and if it is less then it gives the message for supply of water. Water soluble fertilizers are used .After calculating the amount of nutrients required it sets the timer for delivery of fertilizer according to the requirement. The Wifi module is connected to the arduino board for creating the communication. Act as a client that can access the internet by connecting to a router.

[3]IoT Enabled Plant Soil Moisture Monitoring Using Wireless Sensor Networks A.M.Ezhilazhahi1 and P.T.V.Bhuvaneswari,IEEE 2017.

In this paper[3], implementation of the system is done by creating a greenhouse automatic control system based on a wireless sensor network to monitor the indoor conditions. Based on the information collected, the indoor conditions are controlled and monitoring of the crops is carried out which secure the crop from blight and harmful insects. The data gathered is stored either in a database or in a server which is monitored by the user remotely. They monitored the growth of sweet potatoes under a controlled and exposed soil environment. According to limitations in parameter value, greenhouse setup is monitored via relay switches connected to Arduino based embedded units. They have created a temperature and humidity sensor that is placed on plant species. The data from these sensors are gathered continuously and stored in atos pc software which is open source. Then it is uploaded to the server through pc server for remote monitoring.

[4]. Soil Analysis and Crop Fertility Prediction, Komal Abhang Surabhi Chaughule, Pranali Chavan, Shraddha Ganjave, 2018.

In this paper[4], it uses a pH meter to detect the pH content of soil this value is used to estimate the N,P,K values , then this value is used to determine the fertility of the soil.For this testing of various soils was done . This value is inserted into the software then the comparison is done with the database using the classification algorithm.Based on the classification it will give the list of suitable crops for that particular soil.

3. SYSTEM ARCHITECTURE

A system architecture is an abstract architecture,to define a solution based on the concepts. Its main focus is to achieve a life cycle. It helps to describe the entire system.

3.1 Block Diagram

The block diagram is given in Figure 1. Each block is described in this following Section.

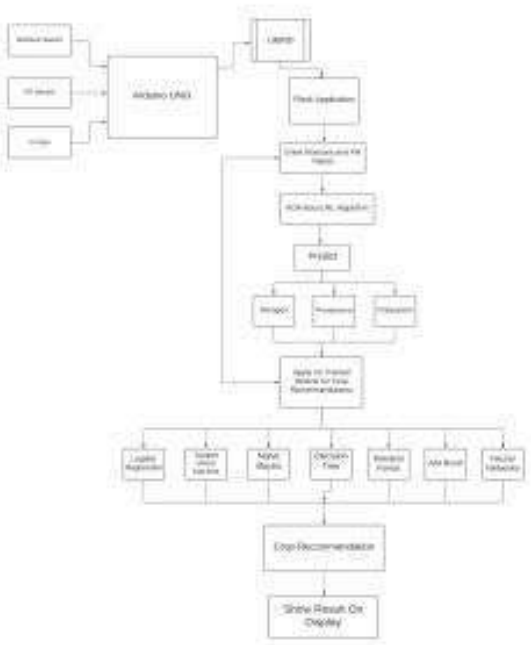


Fig. 1 Block Diagram of Proposed System

A. DESCRIPTION:

In the Project using different sensors through which content of nutrients will be detected and then compared with a dataset through machine learning using classification algorithm which will then give output to the user which will tell the need of fertilizer required for the crop.

Sensor Circuit: This module includes two different types of sensors: pH sensor and moisture sensor. All these sensors are placed inside the pot in such a manner that all the information related to plants such as the moisture content of the soil and the pH can be taken accurately.

pH Meter: The simple method of measuring soil pH by soil meter. Soil pH helps the farmers to get the right value of the pH soil using a pH meter. This tool helps the farmers to get the perfect value of the soil pH. Though the knowledge of the soil acidity is very helpful for the purpose of agriculture.

Moisture Sensor: Moisture sensor helps to sense the moisture of the soil.with this tool the farmers easily get the value of soil moisture.

Flask Application: It is a User Interface Design. used this application to create a website.

Logic Regression: Logistic regression is a supervised classification algorithm used to anticipate the probability of a target variable. It is the simplest ML algorithm that can be used for various classification problems such as spam detection, diabetes detection etc.

Support Vector Machine: In (SVM) support vector machines are supervised learning algorithm models with associated learning algorithms that examine data for classification and regression analysis.

Naive Bayes: Naive bayes is an easy machine learning algorithm based on the bayes theorem, used in a wide variety of classification tasks.It is not a single algorithm but a family of algorithms where all can share a common principle.

Decision Tree: Decision tree algorithm is a member of the family of supervised learning algorithms. The motive of using a decision tree is to generate a training model that can be used to anticipate the class or value of the target variable by learning simple decision rules inferred from training data. It is an illustration representation for getting all the possible solutions to problems and decisions based on given conditions.

Random Forest: Random forest is a machine learning algorithm which helps to solve the both classification and regression problems. It predicts data with high accuracy.

Ada Boost: Ada boost technique follows a decision tree model with a depth equal to one. Ada boost is nothing but the forest of stumps rather than trees. Ada boost algorithm is developed to solve both classification and regression problems.

Neural Networks: A neural network is an algorithm that attempts to identify underlying relationships in a set of data through a process that copies the way the

human brain operates.It helps to understand the impact of increasing and decreasing the dataset vertically or horizontally on a computational line.

Graphical Representation:

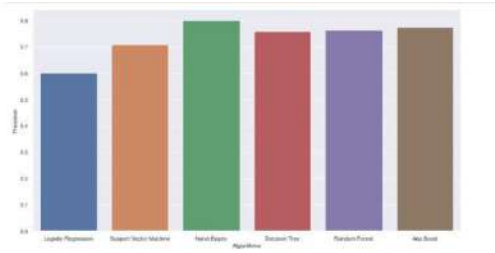


Fig2. Precision Graph

Fig 2.Shows the precision quantifies the number of class predictions that actually belong to the positive class.precision also called positive predictive value.The ratio of correct positive prediction to the total predicted positives.

Precision =TP/TP+FP

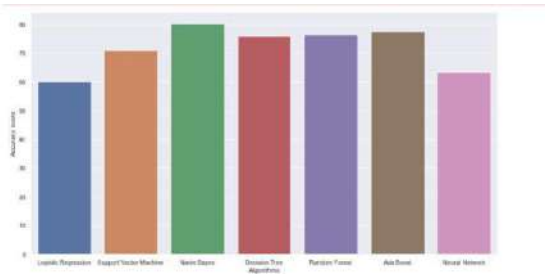


Fig 3. Accuracy Graph

Fig 3. Shows the accuracy is the most intuitive performance measure and it is simply ratio of correctly predicted observations to the total observation.

Accuracy =TP+TN/TP+FP+FN+TN

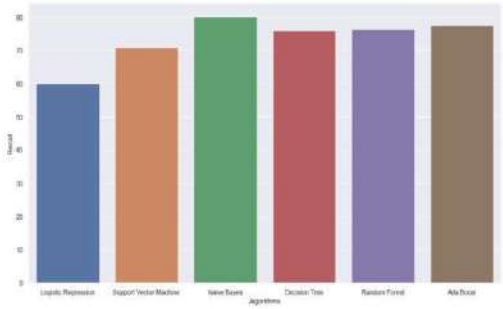


Fig 4. Recall Graph

Fig 4. Shows recall is calculated as the number of true positive and false negative.

Recall = TP/ TP+FN

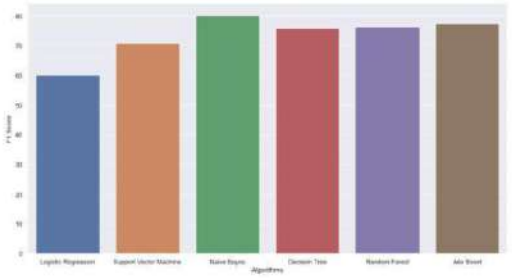


Fig 5. F1 Score Graph

Fig.5 shows F1 score conveys balance between precision and recall.

F1 Score = 2*(Recall* Precision)/(Recall+Precision)

Website Display: Fig 6.Website is to display the values of pH and moisture as well as the N, P, K Values. Data stored in the cloud that can be read on the website display. Website is help to display the output.

Fig 6. Display of Website

REFERENCES

1. Marianah Masrie*, Mohamad Syamim Aizuddin Rosman, Rosidah Sam and Zuriat Janin in “Detection of Nitrogen, Phosphorus,

and Potassium (NPK) nutrients of soil using Optical Transducer" ,2017

2. Cyril Joseph, I Thirunavukkarasu, Aadesh Bhaskar, Anish Penujuru,IEEE 2017.

3. IoT Enabled Plant Soil Moisture Monitoring Using Wireless Sensor Networks. A.M.Ezhilazhahi1 and P.T.V.Bhuvaneswari,IEEE 2017.

4. Komal Abhang Surabhi Chaughule,Pranali Chavan,Shraddha Ganjave,2018. Soil Analysis and Crop Fertility Prediction.

Sentiment Analysis Using Hybrid Feature Extraction For Hotel Reviews

Shivam Naik[1], Akshay Sawant[2], Swapnil Gawade[3] and Dr. Madhu Nashipudimath[4]
Pillai College of Engineering
Department of Information Technology, Mumbai University
New Panvel, Maharashtra (India)
[1]naikseit16e@student.mes.ac.in
[2]sawantar17@student.mes.ac.in
[3]gawadeswasu17ite@student.mes.ac.in
[4]madhumn@mes.ac.in

***Abstract*— Social Networking sites have become popular and common places for sharing wide range of emotions through short texts. These emotions include happiness, sadness, anxiety, fear, etc. Analyzing short texts helps in identifying the sentiment expressed by the crowd. Sentiment Analysis on Hotel reviews identifies the overall sentimentor opinion expressed by a reviewer towards a hotel. Many researchers are working on pruning the sentiment analysis model that clearly identifies and distinguishes between a positive review and a negative review. In the proposed work, we show that the use of Hybrid features obtained by concatenating Machine Learning features (TF-IDF) with Lexicon features (TextBlob) gives better results both in terms of accuracy and complexity. The proposed model clearly differentiates between a positive review and negative review. Since understanding the context of the reviews plays an important role in classification, using Hybrid features, helps in capturing the context of the Hotel reviews and hence increases the accuracy of classification.**

***Keywords*— Sentiment Analysis, Classification, Feature Extraction, Machine Learning, Naïve Bayes, Natural Language Processing**

I. INTRODUCTION

In this modern age where the internet is growing rapidly, the existence of the internet can make it easier for tourist to find any information. In the field of tourism hotel, internet is very helpful in promotion of hotel. Tourists usually tell the experience during the hotel by writing reviews on the internet. Hence many hotel's reviews are found on the internet. The impact on hotel owners is that they can take advantage of reviews on the internet to improve and evaluate their hotels. With the availability of reviews on the internet with large numbers, tourists can't understand all the reviews they read whether they contain positive or negative opinions. It takes a sentiment analysis to quickly detect if the review is a positive or negative.

Sentiment analysis is the interpretation and classification of emotions (positive, negative and neutral) within text data using text analysis techniques. Sentiment analysis allows businesses to identify customer sentiment towards products, brands or services in online conversations and feedback. For a hotel business, reviews about various aspects like Maintenance, Food, Hospitality, Room Neatness, Response from the staff of the hotel etc. plays a major role for recommender system. The Customer's sentiments regarding to a hotel depends upon the facilities he/she got from that hotel like cleanliness, location of the hotel, services provided by the hotel like free wi-fi, multilingual staff, bar/lounge, babysitting rooms, wheel chair etc. The sentiments can be expressed in the form of excellent, good, average, poor, terrible etc. Generally, customers want to express their feelings also with these rating and review values. Sentiment analysis, however helps businesses make sense of all this unstructured text by automatically tagging it.

The importance of online reviews plays a vital role: it is the key to your hotel standing on the online portals which in turn leads to greater business and increased revenue outcomes. Precise management for your brand on online portals will reassure potential customers and motivate them to opt for the hotel without a second thought in their mind. Getting the reviews classified to gain insights from it, is now an important part of the hotel business. Reviews tell the story of how the customer feels about the services which the hotel is providing. The positive reviews can also be used to promote the good efforts of the hotel just as important as to take the negative reviews into account. Sentiment analysis helps to improve the hotel business in several ways, from preventing a shrinking reputation in the market to understanding how the guests feel about their facility.

Since there are tons of reviews available through different online platforms analyzing by themselves is no longer accountable for hotel businesses. They require accurate, reliable, fast, and efficient automated systems that can provide better findings to empower business decisions. Sentiment analysis is indeed required to automate the process of determining whether a review expresses a positive, negative, or neutral opinion about the hotel and its services. With the help of sentiment analysis, hotels can save limitless time labeling customer data such as reviews, ratings, and comments on social media platforms. Sentiment analysis is required by the hotels to monitor their brand value on online portals, and gain information from customer feedback, and in turn, apply them to improve themselves.

The main objective of this project is to develop a model which identifies the overall sentimentor opinion expressed by a reviewer towards a hotel. Reviews are short texts that generally express an opinion about hotels or products. These reviews play a vital role in the sales. People generally look into review sites like Trip Advisor to know about hotel, location, review and ratings. Hence it is not

only the Word of Mouth; reviews also play a prominent role in this regard. In other words, Sentiment Analysis on hotel reviews makes the task of Opinion Summarization easier by extracting the sentiment expressed by the reviewer.

The contents of the paper are divided into five sections. Section II presents an overview of the literature survey of previous works on sentiment analysis. Section III presents the methodology. Section IV shows experimental results and the paper is concluded in section V.

II. Related Work

Earlier works on sentiment classification using machine learning approaches were carried by Pang et al. in 2002 [8] on movie reviews using n-gram approaches and Bag of Words (BOW) as features and model were trained using different classifiers. Similar work was done by Tripathy et al [9] where TF, TF-IDF was used for the conversion of the text file to a numerical vector. Experimentation was done with n-gram approaches and it's combination to get the best results. Apart from using statistical features use of HFEM made the model more efficient in terms of accurate classification by adding the advantages of individual feature extraction method. Results obtained were highly promising both in terms of space complexity and classification accuracy.

H. M. Keerthi Kumar, B. S. Harish, H. K. Darshan [1] proposed sentiment analysis model employed on IMDb Movie Reviews. The input for the proposed model was the set of reviews whose polarity was to be determined. The task was carried out in the following phases: preprocessing the dataset, Feature Extraction (Both Statistical and Lexicon approach), feature selection and finally classification using hybrid features. The proposed work captured the polarity of a word and determined how important the word is for the classification task. The capturing phase was done through the features generated using Hybrid Feature Extraction Method (HFEM). In addition, various feature selection methods such as Chi-Square, Correlation, Information Gain and Regularized Locality Preserving Indexing (RLPI) were applied for the features extracted by statistical methods. The Lexicon based feature extraction method extract features based on the Lexicon dictionaries. Features from both methods were combined to form a new feature set which is of lower dimension when compared to the initial dimension of the input space. The new features set was classified using various classifiers such as Support Vector Machines (SVM), Naïve Bayes (NB), K- Nearest Neighbor (KNN) and Maximum Entropy (ME) on IMDb movie review dataset.

Sentiment reviews classification using Hybrid Feature Selection (SRCHFS) proposed by K.Bhuvaneshwari, R.Parimala [2] extracted synsets feature set coupled with correlation feature selection method can improve the perfomance of sentiment classification. A set of cognitive synsets is selected using WordNet based POS (Part of Speech). Support Vector Machine (SVM) classifier was used for sentiment classification on a data set of movie reviews, Multi Domain product reviews, Amazon Cell phone reviews and Yelp Restaurant reviews. The confusion matrix is used to evaluate the performance of sentiment classification.

Indrajeet Kaur Chhabra, Gend Lal Prajapati [3] proposed Sentiment analysis of amazon canon camera review using Hybrid method. The Lexicon based approach used the built-in dictionaries of words and their semantic orientation to find the polarity of customer product review. In this work, the SentiWordNet lexicon resource is used to find the positive and negative polarity of words. A hybrid approach is used in which the results of lexicon based approach, the positive and negative polarity of the words is used to train the classifier. A support vector machine is used as it is a binary classifier so best worked for classification of reviews into positive and negative class. The experiments show that the proposed hybrid approach performs better than the machine leaning approach and lexicon-based approach.

The three major techniques Statistical methods, Knowledge-based methods and Hybrid techniques were used in Sentimental analysis Model of Hotel review from TripAdvisor created by Vaibhav Singh, Aayushi Mahajan, Deepanshi Chaudhary [4]. The goal of Knowledge based methods is to extract knowledge by classifying text based on categories explicitly present in words such as awesome, sad, happy, unfortunate, and poor etc. The hybrid approach as the name suggests it constitutes both the above methods i.e. the statistical learning approach and the knowledge-based method for calculating the polarity scores. The reason for combining is to gain high accuracy and stability of the system at the same time. The TextBlob library from python is widely used for sentiment analysis and is built on the top of NLTK. The Text blob sentiment calculates the sentiment polarity and subjectivity. We evaluated the analysis system on a corpus of 738 hotel reviews crawled from the web. The results were as out of 738 reviews we found that 97.3 percent of the total reviews were listed as positive, 2.6 percent of the total reviews were listed as negative and 0.1 percent of the reviews are neutral.

Venkateswarlu Bonta, Nandhini Kumaresh and N. Janardhan [5] used NLTK, Text blob and VADER Sentiment analysis tool to classify the movie reviews and made a comparison on these tools to find the efficient one for sentiment classification. Arif Abdurrahman Farisi, Yuliant Sibaroni and Said Al Faraby [6] experimented the results using pre-processing and feature selection with 10 fold cross validation which gave an average F1-score of more than 91%. Maximum Entropy classifier does not assume independence of features so theoretically it may outperform Naive Bayes. Also the algorithm is more difficult and the learning process is slower as well. To improve the evaluation accuracy, Vikas Malik, Amit Kumar [7] build an LSTM network, and benchmark its results compared to the NLTK machine learning implementation. The content-based recommendation system implemented by P. Sanjay Bhargav, G. Nagarjuna Reddy, R.V. Ravi Chand, K.Pujitha, Anjali Mathur [10] implied matching of attributes from a user profile in which preferences and interest are stored with attributes of content object. If a string is found in both the profile and the document, a match is made and the document is considered as relevant. Melville et al. [11], worked on extracting features using lexicon methods. Positive and negative word counts present in the text were used as the background lexicon knowledge and then the probability that a document belongs to a particular class was calculated. Use of pooling multinomial classifiers which incorporate both training examples and the background knowledge was the major contribution.

III. METHODOLOGY

In the proposed work, Sentiment Analysis using Hybrid Feature Extraction for Hotel Reviews is performed. The task of sentiment analysis is carried out in the following phases: preprocessing the dataset, feature extraction (Both Statistical and Lexicon methods), feature selection and finally classification using Naïve Bayes Classifier. A comparative analysis of the accuracies obtained by 3 feature sets (Lexicon, TF-IDF, Hybrid features) is also shown.

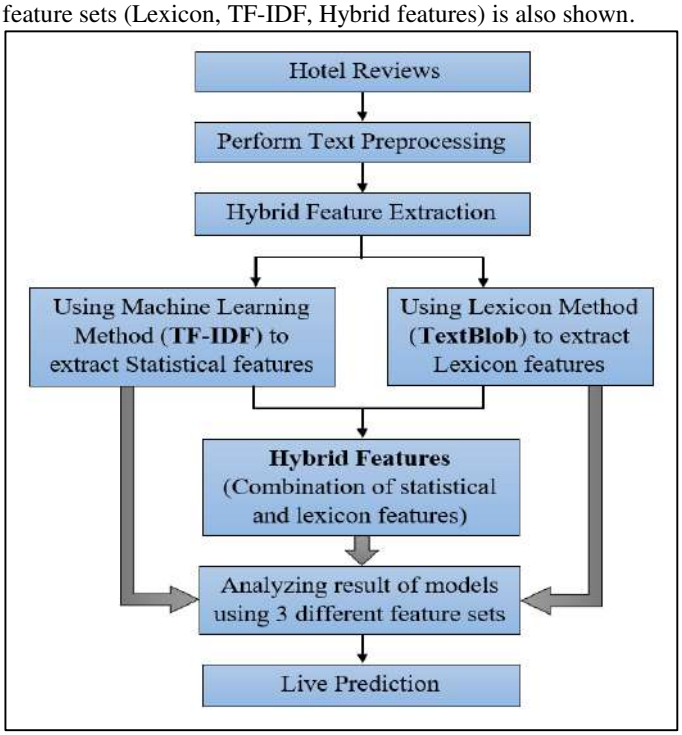


Fig 1. Proposed System Architecture for Hotel reviews

1. Text Preprocessing

Data Preprocessing is a process for making low quality data into high quality data, making it easy to process. The raw data having polarity is highly susceptible to inconsistency and redundancy. Preprocessing includes :

a) Removal of all punctuations, numbers, symbols - The reviews which need to be analyzed consist of numbers, symbols and punctuations which does not influence on sentiment analysis because of its neutral polarity. Hence they are removed.
b) Case folding - It is the process whereby all the letters are converted into lowercase
c) Stopword removal - Stopword removal is the process of removing less important words that often appear on documents. To shorten the classification process, it can eliminate stop words such as "is", "the", "and" etc.
d) Tokenization - It is a way of separating a piece of text into smaller units called tokens. Here, tokens can be either words, characters, or subwords.
e) Lemmatization - in lemmatization, we try to reduce a given word to its root word with the help of vocabulary. The root word is called a stem in the stemming process, and it is called a lemma in the lemmatization process. For example, words like 'studying', 'studied' will reduce to it's root word 'study'.

Further, the preprocessed data is used for feature extraction in the next phase.

2. Feature Extraction

Feature Extraction is the process of extracting relevant features. In this work, feature extraction is carried in two different parallel stages namely- Machine learning based feature extraction and Lexicon based feature extraction.

a. Machine learning based feature extraction method is used to extract the features using popularly known technique Bag of Words, wherein the column corresponds to words and row corresponds to value of weighing measures such as Term Frequency (TF), Document Frequency (DF), Term Frequency-Inverse Document Frequency (TF-IDF).

- Term Frequency (TF) = No of times word occurs in review / Total no of words in review. TF says what is probability of finding a word in a document. More often word occurs in review, higher will be the term frequency.
- Document Frequency (DF) = No of documents in which word is present. We consider one occurrence if the term consists in the document at least once, we do not need to know number of times the term is present.
- Inverse Document Frequency (IDF) = No of documents / Document Frequency. IDF is the inverse of the document frequency which measures the informativeness of word. When we calculate IDF, it will be very low for the most occurring words in the corpus. Also IDF value will be high for rare words in the corpus. This finally gives what we want a relative weightage. But there are few problems with the IDF, in case of a large corpus, say 100,000,000 , IDF value explodes , to avoid it we take the log of idf . Also during query time, when a word which is not in vocab occurs, the DF will be 0. As we cannot divide by 0, we smoothen the value by adding 1 to the denominator.
- Term Frequency-Inverse Document Frequency (TF-IDF) : TF-IDF(word,review) = TF(word, review) * log(No of reviews/(DF + 1)). It is a the right measure to evaluate how important a word is to a document in a corpus.

b. Lexicon based feature extraction method used in the this work extracts features using TextBlob. It uses sentiment lexicon with information about which words and phrases are positive and which are negative.

- TextBlob : Textblob is a python library for processing textual data. It provides a consistent API for common natural language processing (NLP) tasks. Textblob is a sentence level analysis. First, it takes a dataset as the input then it splits the review into sentences. A common way of determining polarity for an entire dataset is to count the number of positive and negative sentences/reviews and decide whether the response is positive and negative based on total number of positive and negative reviews. Polarity and subjectivity of a given review can be known using sentiment() function. It returns a named tuple with two parameters called polarity and subjectivity. The polarity score is ranging from -1 to 1 and subjectivity ranges are from 0 to 1 where 0 is most objective and 1 is most subjective.
  Example: Review = Textblob ("the movie was interesting.")
  review.sentiment - Sentiment(polarity=0.5, subjectivity=0.5)

Combination of features extracted through different feature extraction methods will increase the overall performance of the

model. Combining Machine learning based features and Lexicon based features helps in identifying the overall polarity of the review more accurately.

3. Classification

Naive Bayes Algorithm : Naive Bayes is a simple probabilistic machine learning algorithm based on Bayes theorem with the independence assumptions between features. 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 it often outperforms more sophisticated classification methods.

Naive Bayes Formulation :

$$P(c|x) = P(x|c).P(c) / P(x)$$

- P(c|x) is the posterior probability of class (target) given predictor (feature).
- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

Naive Bayes works best for text classification problems as it has a higher success rate than other algorithms. It is a benchmark against other algorithms for comparison in perfomance. Multinomial Navie Bayes algorithm is used for classification provided by Scikit-learn library.

IV. Experimental Setup

A] Dataset

Most of the hotels ask reviews from the customers, based on that customer satisfaction improves. So reviews plays a vital role for the successful growth of the Hotel. Hotel-reviews dataset is used in the proposed work. The dataset consists of 1000 rows and 2 columns. Due to the limitations of computational resources, small balanced dataset is used.

B] Software and Hardware specifications

We have used Windows 10 operating system, Python programming language and also Jupyter Notebook editor. The hardware used for this project is intel i3 processor, 512 GB HDD, 8 GB RAM

C] Experimentation

The supervised machine learning model is built using training data (which has input as well as output). Prediction is made on the test data (unseen data which does not have an output label) using the same model. To know the effectiveness of the model, there are some measures that will evaluate the performance of the model on the test dataset. There are many performance metrics such as accuracy, precision, recall, F1-score, ROC curve, etc. each having its advantages and disadvantages. In this work, accuracy is used as performance metric.

Accuracy : It is the most intuitive performance measure. It is the ratio of correctly classified points (prediction) to the total number of predictions. It works well only if there are equal number of samples belonging to each class. Its value ranges between 0 and 1. To calculate the accuracy, Python's metrics.accuracy_score module is used.

D] Result Analysis

Using three feature sets i.e Statistical features(tf-idf), Lexicon features(TextBlob), and Hybrid features models are build and performance is evaluated for each to compare results obtained by 3 different models on basis of accuracy as performance measure.

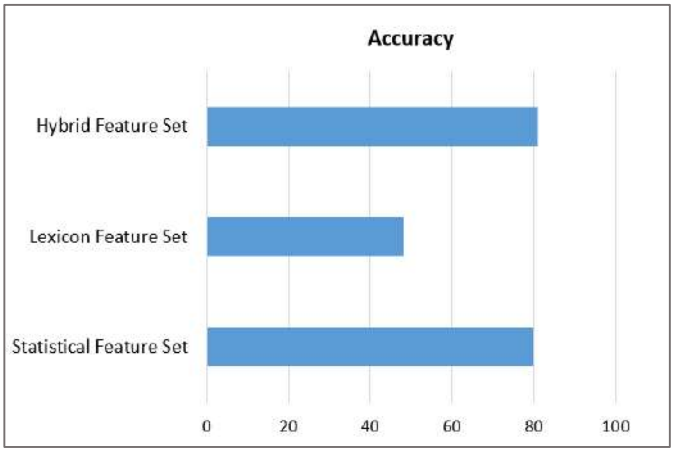


Fig 2. Result Analysis

The results shown in Table 1 proves that the accuracy of hybrid approach is better than the machine leaning approach or the lexicon approach alone.By comparing the perfomance of models which used 3 different feature sets, results show that model using hybird feature set is the most optimal with an accuracy of 81%. Model using statistical feature set also performed well but by concatenating lexicon features to it made a slight improvement in overall perfomance. Also the Naive Bayes classifier correctly classified the user input review.

E] Output Block Description

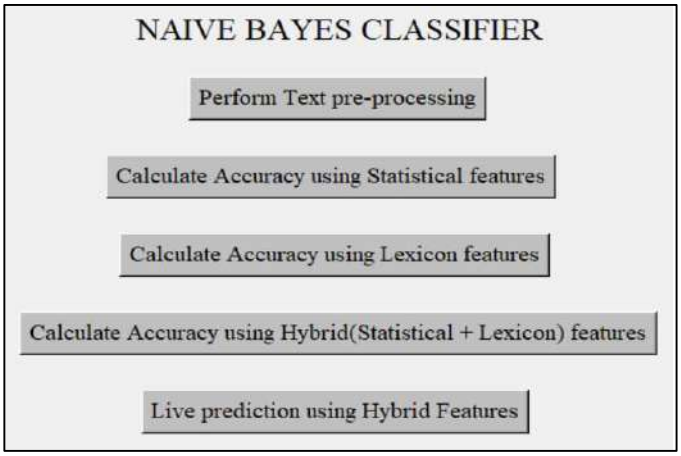


Fig 3. GUI for user input

Firstly text preprocessing is performed to clean the data. Accuracy using Statistical, Lexicon and Hybrid Features is calculated seperately. Lastly live prediction is shown where-in user enters a review and the model says whether that review is positive or negative.

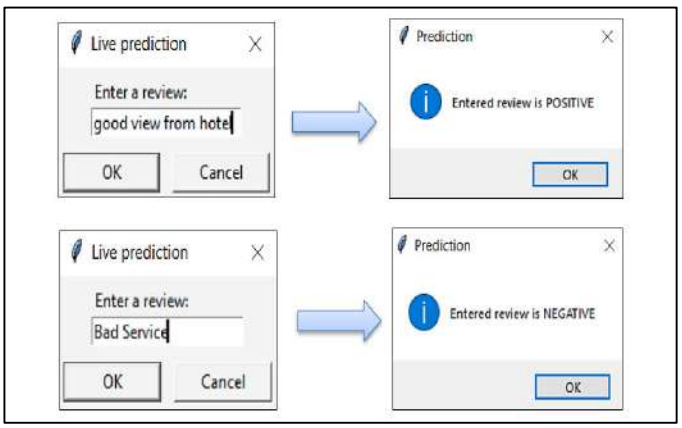


Fig 4. Live Prediction

Live prediction window clearly differentiates between a positive review and negative review given any user input review.

V. CONCLUSION

In this paper, Hybrid Feature Extraction Method (HFEM) is used to extract features from machine learning and lexicon based feature extraction methods. Initially machine learning features (TF-IDF) are high dimensional in nature. The top features are selected using the corresponding tfidf value. On the other hand lexicon features are extracted using TextBlob. It uses sentiment lexicon with information about, which words and phrases are positive and which are negative.The polarity score ranges from -1 to 1 and subjectivity score ranges from 0 to 1 where 0 is most objective and 1 is most subjective. Combining Machine learning based features and Lexicon based features helps in identifying the overall polarity of the review more accurately. To demonstrate the effectiveness of the proposed work, we used Naïve Bayes classifier on Hotel review dataset. Use of Hybrid Feature Extraction Method (HFEM) makes the model more efficient in terms of accurate classification by adding the advantages of individual feature extraction method. HFEM improves the space complexity by reducing the input space to minimal number of features that are sufficient to represent the review content. Thus, results obtained are highly promising both in terms of space complexity and classification accuracy. In future work, we will include more lexicon features to the feature subset and thereby expect to increase the classification accuracy.

ACKNOWLEDGMENT

We would also like to thank our mentor Dr. Madhu Nashipudimath who helped us regarding any and all queries while working on the project.We would like to thank everyone who provided us an opportunity to work on projects like this thereby increasing our knowledge. We would like to thank our Principal Dr. Sandeep Joshi who always encouraged and motivated us. We would also like to express our gratitude to our H.O.D of Information Technology Department Dr. Satishkumar Varma for giving us this opportunity and for motivating us to do innovative things that will be beneficial for our future.

REFERENCES

[1] H. M. Keerthi Kumar, B. S. Harish, H. K. Darshan. **"Sentiment Analysis on IMDb Movie Reviews Using Hybrid Feature Extraction Method."** International Journal of Interactive Multimedia & Artificial Intelligence (2019).

[2] K. Bhuvaneswari, R. Parimala. **"Sentiment Reviews Classification using Hybrid Feature Selection"** International Journal of Database Theory and Application Vol.10, No.7 (2017)

[3] Vaibhav Singh, Aayushi Mahajan, Deepanshi Chaudhary. **"Sentiment Analysis of Hotel Reviews from TripAdvisor"** International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 07 Issue: 06 | June 2020

[4] Indrajeet Kaur Chhabra, Gend Lal Prajapati. **"Sentiment Analysis of Amazon Canon Camera Review using Hybrid Method"** International Journal of Computer Applications Volume 182 – No.5, July 2018

[5] Venkateswarlu Bonta, Nandhini Kumaresh and N. Janardhan. **"Comprehensive Study on Lexicon Based Approaches for Sentiment Analysis"** Asian Journal of Computer Science and Technology ISSN: 2249-0701 Vol.8 No.S2, 2019, pp. 1-6

[6] Arif Abdurrahman Farisi, Yuliant Sibaroni and Said Al Faraby. **"Sentiment Analysis on Hotel Reviews using Multinomial Naïve Bayes classifier"** IOP Conf. Series : Journal of Physics (2019)

[7] Vikas Malik, Amit Kumar. **"Sentiment Analysis of Twitter Data Using Naive Bayes Algorithm"** International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-169 Volume: 6 Issue: 4 (2018)

[8] B. Pang, L. Lee, and S. Vaithyanathan. **"Thumbs up?: sentiment classification using machine learning techniques."** In Proceedings of the ACL-02 conference on Empirical methods in natural language processing- Volume 10, Association for Computational Linguistics, pp. 79-86. 2002.

[9] A. Tripathy, A. Agrawal, and S.K. Rath. **"Classification of sentiment reviews using n-gram machine learning approach."** Expert Systems with Applications, Vol. 57, pp. 117-126. 2016.

[10] P. Sanjay Bhargav, G. Nagarjuna Reddy, R.V. Ravi Chand, K.Pujitha, Anjali Mathur **"Sentiment Analysis for Hotel Rating using Machine Learning Algorithms"** International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-6, April 2019

[11] P. Melville, W. Gryc, and R. D. Lawrence. **"Sentiment analysis of blogs by combining lexical knowledge with text classification."** In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp. 1275-1284. 2009.

[12] H X Shi and X J Li **"A sentiment analysis model for hotel reviews based on supervised learning"** in in International Conference on Machine Learning and Cybernetics China (2011)

[13] T Ghorpade and L Ragha **"Featured Based Sentiment Classification for Hotel Reviews using NLP and Bayesian Classification"** in International Conference on Communication, Information & Computing Technology (ICCICT) Mumbai India (2012)

[14] Wararat Songpan "**The Analysis and Prediction of Customer Review Rating Using Opinion Mining**" 2017 IEEE 15th International Conference on Software Engineering Research, Mangement and Applications (SERA)

[15] A. Ortigosa, J. M. Martín, and R. M. Carro. **"Sentiment analysis in Facebook and its application to e-learning."** Computers in Human Behavior Vol. 31, pp.527-541. 2014.

[16] L. Zheng, H. Wang, and S. Gao. "**Sentimental feature selection for sentiment analysis of Chinese online reviews**." International journal of machine learning and cybernetics, Vol. 9, no. 1, pp.75-84. 2018.

FAKE PROFILE DETECTION USING DEEP LEARNING

Yadnika Birari
Pillai College Of Engineering
Panvel, Maharashtra
yadnikabirari99@gmail.com

Abhishek Chaudhuri
Pillai College Of Engineering
Panvel, Maharashtra
abhichou19@gmail.com

Sanjana Darne
Pillai College Of Engineering
Panvel, Maharashtra
sanjanadarne01@gmail.com

Prof.Madhura Vyawahare
Pillai College Of Engineering
Panvel, Maharashtra
madhuravyawahare@mes.ac.in

***Abstract*—These days each and every person has access to the internet, this means that most of the internet users in today's date will be unable to differentiate between what's safe and what's threatening to them. As the number of internet users are increasing day by day the users of OSN (online social networks) are also increasing which is directly proportional to the increase in all kinds of fake and malicious attacks on the users of these online social networks. On top of that the open nature of these online social networks have made them vulnerable to various attacks including the sybil attacks. As the online social platforms are growing more and more popular the identity clone attacks that aim at creating fake identities for malicious purposes are also growing directly proportional to it. In this system we will make the use of deep learning to check if the twitter id provided to the system is fake or genuine. And for doing that we will make the use of RNN LSTM in deep learning and the string comparators for the comparison of the two different strings.**

i.Keywords*: *OSN(online social networks), sybil attacks, fake identities, twitter, string comparators, Deep Learning.

I. Introduction

In this era the online social networks are considered to be the most popular platforms on the internet. It plays a major role for the users of the internet to perform their everyday actions such as news reading, content sharing, messages posting, product reviews and event discussions etc. The massive amounts of personal data of the users coupled with the open nature of these online social networks have made these online social networks vulnerable to various attacks including the sybil attacks. As the online social networks are becoming increasingly popular the identity clone attacks that aim at creating fake identities for malicious purposes are also becoming a growing concern these days. There are multiple types of spammers that coexist in the online social networks.

We are here proposing a system that can be used to detect the fake profiles present on the online social network (Twitter). This system will use deep learning to generate a base tweet and then using the string comparators it will compare the different tweets and in the end we will have the results if the Twitter id is genuine or it's fake. Thus this system will be of great use for the people as well as the host of the social media service.

II. Literature Survey

The authors Sarah Khaled, Hoda M. O. Mokhtar came up with solving the problems on fake profile detection in social media platforms. This particular approach of identifying fake social media profiles was classified into the approaches aimed at analysis of the individual account and the approaches capturing the activities spanning in a large sample of accounts. The classification of these profiles based on their features made the use of several machine learning algorithms. [1]. Binghui Wang, Le Zhang proposed a system called SybilBlind which is a structure-based framework that is used to detect sybils in the social media platforms without a manually labeled training set. The evaluation is that the SybilBlind both theoretically and empirically, as well as compared it with Sybil detection methods that we adapt to detect Sybils when no manually labeled training sets are available. Their empirical results demonstrated the superiority of SybilBlind over the adapted methods. [2]. Mohammadreza M, Mohammad Eb. has proposed a model that makes the use of a resampling approach. The resampling approach means changing the distribution of training sampling sets by making the required changes to the data that is by processing the data. Balancing the datasets is one of the approaches used towards improving the class efficiency. This particular system also showed the use of principal component analysis. The basic idea of principal component analysis (PCA) is one of the multivariate classical methods and perhaps the most ancient and most popular one. Mostly all these machine learning methods train the classifiers using the machine learning algorithms. Attribute similarity, network friend similarity and IP address analysis are some of the social network attributes on which the classifiers are based[3]. On the same line we have developed a system to give better performance for detecting fake profiles.

III. Algorithms

In order to identify the fake profiles, the Recurrent Neural Networks along with its various algorithms have been implemented to compare the strings and their values have been calculated in mean, algorithms used are: the levenshtein distance, dice's coefficient and Long short term memory (LSTM).

Recurrent neural networks:
As humans cannot understand the meaning of a word or sentence without its previous information. Every sentence is to be linked with the previous sentence in order to understand the full text. Similarly, the traditional neural networks have been facing issues and are unable to act like this. This is where the Recurrent Neural Networks helps and addresses the value. This type of network is with loops which allows the data to persist.

Long Short Term Memory(LSTM):
LSTM networks are a type of neural network which is capable of learning order dependencies in sequence of prediction problems. This usually helps in big problem domains such as machine translations, speech recognitions and many more complex issues. LTSMs is the complex area of deep learning. LSTMs are designed to avoid the occuring long-term dependency problems and remembering information is in their behavior and it does have to struggle to learn.

The Levenshtein distance:
It is a string metric for measuring the differences between strings having two sequences. The insertion, deletion or substitution between the two words is calculated using levenshtein distance which is required to change one word into the other. It may also refer for editing distance even if it also denotes a large cluster of distance metrics. It is generally related to pairing wise string alignments. It uses:

- a single distance vector instead of using a matrix.
- a loop unrolling on the loop on its outer side.
- by removing common prefixes and postfixes.
- minimizing the comparisons.

Dice's coefficient:
It measures how similar two sets are with each other. Here, it can be used to check the similarities between two strings in terms of the number of similar bigrams that is the pair of adjacent letters in a string.

IV. PROPOSED SYSTEM

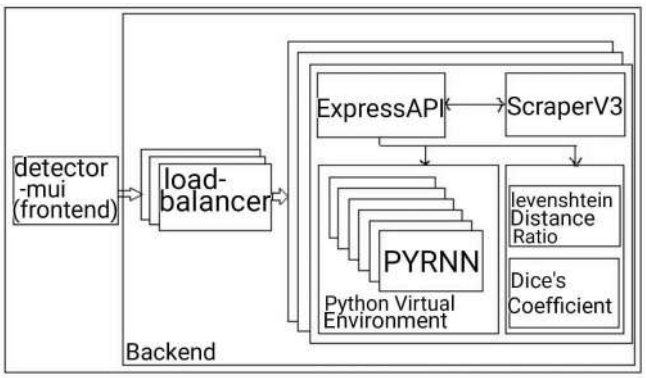


Fig.1. Flow of the project

The above mentioned system has three phases depending on the following:-

Phase1:- Scraping the data
In this section the system scraps the Twitter data using the Twitter metadata api. This is the first phase of the system where the user of the system puts in the Twitter id that has to be analysed and then the system starts scrapping the Twitter id for all possible tweets made by that id in the past few months.

Phase2:- Generating the base tweet
In this section the system creates a child process and the whole text generation process takes place in a Python virtual environment. The text generation that is the generation of the base tweet is done using deep learning. Here we make the use of LSTM RNN and generate a base tweet which has all the possible characteristics of the tweets that we have scrapped in the phase 1 of the system. After generating the base tweet we hop on to the 3rd phase of the system.

Phase 3:- Comparing the strings
In this section the system compares the two different strings in two different ways. First of all we get the base tweet that was generated using deep learning and then compare it with a randomly picked tweet from the dataset of tweets scrapped in phase one. Now the comparison is done using the string comparators and in this system we have made use of two string comparators which are the Levenshtein distance ratio and the Dice's coefficient. At first we perform the Levenshtein distance ratio and find out the random result and the average result then similarly we perform the Dice's coefficient and find out the random result and the average result and in the end to make the result more accurate we take the average of all the four results.

V. IMPLEMENTATION DETAILS AND RESULTS ANALYSIS

For implementation we have used 'svelte application' to show the result in our application based website.

1. Back-end.

   Command to start Back-end by scraping the data from Twitter api in real time by using a scraping algorithm for comparing the tweets.

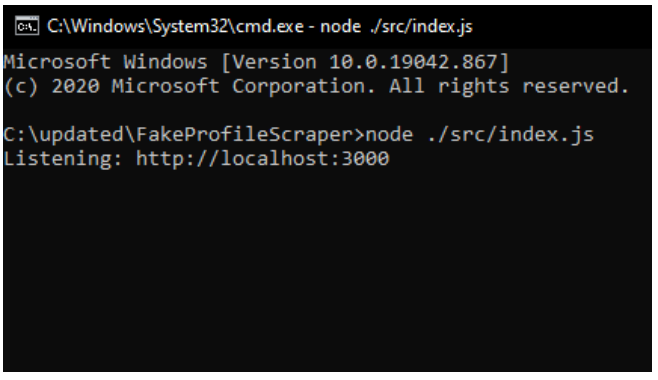


Fig.2. Back-end

2. Front-end.

   Command to start Front-end and use a generated localhost id to start the svelte application.

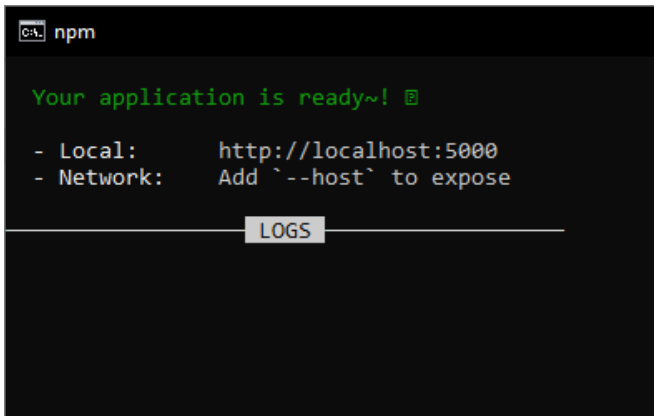


Fig.3. Front-end

3. The home page.

   Svelte application is used to show the results of our application based website which has textbox to enter the twitter username.

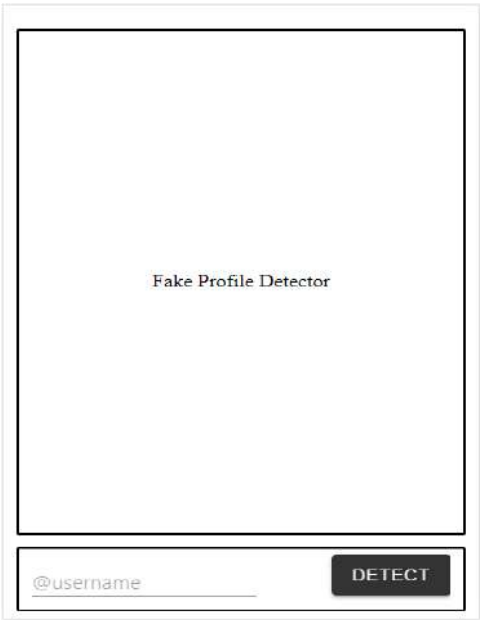


Fig.4. The home page

4. The result of a real account user.

   Entering the username of a real account and the result calculated using the Random selection and mean of the Dice's Coefficient and Levenshtein Distance in percentage. From Figure 5 we can understand that based on the score we can conclude that profile is a real profile. As we can see in the FIgure 5 shows the low percentage when we pass the id "@elonmusk" it shows 37% after calculations.

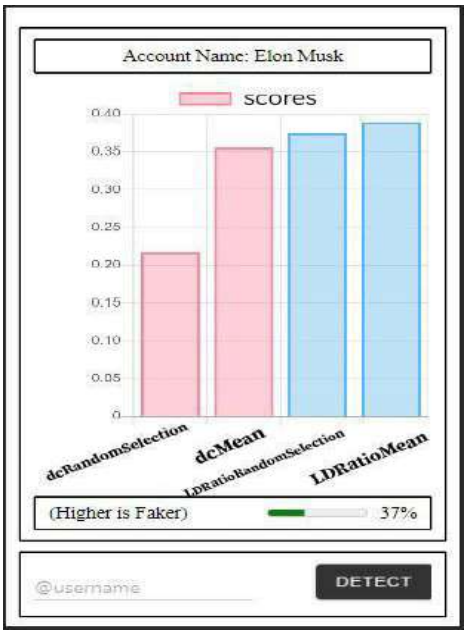


Fig.5. Result of Real Account User

5. The result of a bot account user.

   Entering the username of a bot account and the result calculated using the Random selection and mean of the Dice's Coefficient and Levenshtein Distance in percentage. From Figure 6 we can understand that based on the score we can conclude that the user is bot. As we can see in the FIgure 6 shows the low percentage when we pass the id "@bot_of_jess" it shows 84% after calculations.

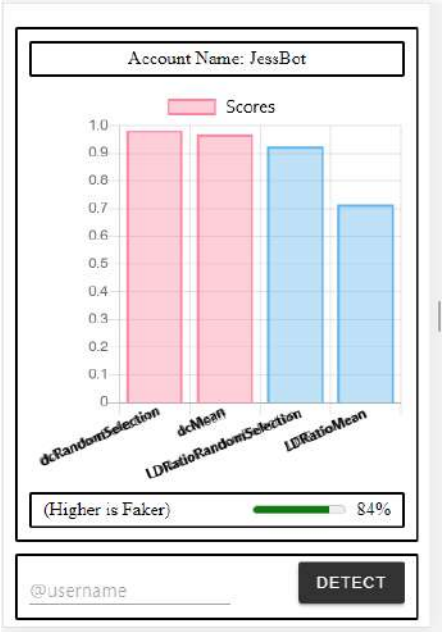


Fig.6. Result of Bot Account User.

6. Result of a user with no tweets.

If the user has not tweeted anything, then the model is capable of displaying the result as 'user has no tweets'. As identifying real, fake or bot accounts is very dependent on tweets posted by users. Identifying this attribute plays a vital role in fake profile detection. Figure 7 shows the result when we pass the id "@YadnikaB" without any tweet.

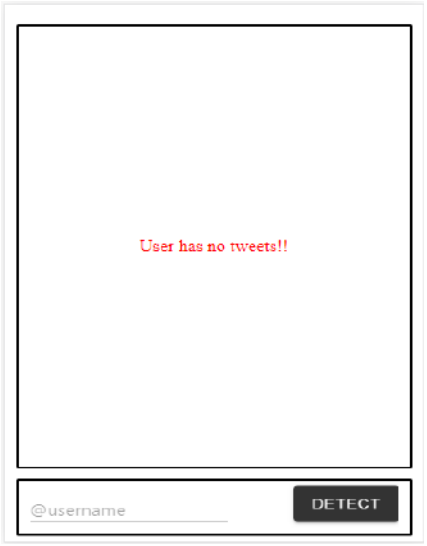


Fig.7. Result of User with no tweets

7. Result of a user if the user account is suspended.

Displaying the result as 'user has been suspended' if the user account is suspended by twitter then it is also identified by our model. As we can see in the figure 8 for user id "@DarneSanjana7" it displays the user is currently suspended.

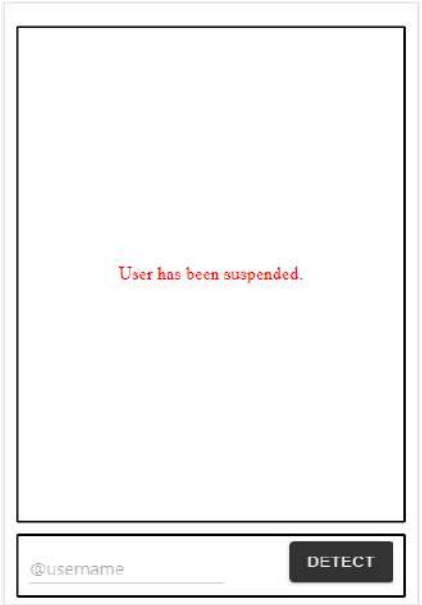


Fig.8. Result of Suspended User Account

VI. Future scope

There are many features that can be included in this project such as:

I. The future work concentrates on replacing more easy algorithms to detect fake user accounts such as replacing LSTM-Rnn with Transformers.
II. To apply the algorithms on different social media platforms to identify fake users.
III. With proper accuracy of the result, this project aims to help cyber security branches.

VII. Conclusion

In this paper, we presented a Fake Profile Detection System using Deep Learning; our project detects the fake or bot twitter profiles by using deep learning algorithms. This project can help the social media platforms hosts as well as the social media users to be protected from all the fake profile related threats.The future scope of the system is to make it more reliable and to include more characteristics to determine the genuineness of the profile.

VIII. References

[1] Yasyn Elyusu, Zakaria Elyusu, and M'hamed Ait Kbir, "Social Networks Fake Profiles Detection Using Machine Learning Algorithms," Faculty of Sciences and Technologies, Tangier, Morocco,In book: Innovations in Smart Cities Applications Edition 3 (pp.30-40), 10.1007/978-3-030-37629-1_3(2019).

[2] Sarah, M. O. Mokhtar,Neamat El-Tazi, "Detecting Fake Accounts on Social Media "Faculty of Computers and Information, Cairo University,Cairo Egypt 2018 IEEE International Conference on Big Data (Big Data), 10.1109/BigData.2018.8621913(2018).

[3] Binghui Wang, Le Zhang, and Neil Zhenqiang Gong ECE , "SybilBlind: Detecting Fake Users in Online Social Networks without Manual Labels", Department, Iowa State University(2018).

[4] Mohammadreza Mohammadrezaei, Mohammad Ebrahim Shiri ,and Amir Masoud Rahmani, "Identifying Fake Accounts on Social Networks Based on Graph Analysis and Classification Algorithms",Computer Science, University of Human Development, Sulaymaniyah, Iraq, Content published prior to 2017 is hosted on the Wiley Online Library, DOI: 10.1155/2037(2017).

[5] Dr. Sanjeev Dhawan, Ekta, "Implications of Various Fake Profile Detection Techniques in Social Networks", UIET, Kurukshetra University, 136119, Kurukshetra, Haryana,India, February 2016IOSR Journal of Computer Engineering 02(02):49-55, 10.9790/0661-15010020249-55 (2016).

[6] Shalinda Adikari and Kaushik Dutta, "Identifying Fake Profiles in LinkedIn", PACIS Proceedings, AISeL, PACIS 2014 Proceedings. 278, https://aisel.aisnet.org/pacis2014/278/ ,(2014).

[7] Dr. S. Kannan, Vairaprakash Gurusamy, "Preprocessing Techniques for Text Mining", UEIS, New Delhi, India, 2019 JETIR May 2019, Volume 6, Issue 5 , www.jetir.org (ISSN-2349-5162) ,(2015).

[8] Thomas, Kurt, Damon McCoy, Chris Grier, Alek Kolcz, and Vern Paxson. "Trafficking Fraudulent Accounts: The Role of the Underground Market in Twitter Spam and Abuse." Presented as part of the 22nd {USENIX} Security Symposium ({USENIX} Security 13), pp. 195-210. (2013).

Journal of Information Technology, Volume 8, Issue 1, 2020 - 2021.

Web Application Penetration Testing Tool

Yugabdh Pashte, Yash Patel, Ruthvik Shetty
Information Technology Department
Pillai College of Engineering, Mumbai University
New Panvel, Raigad, Maharashtra, India
pashteyugpr17ite@student.mes.ac.in
patelyashd17@student.mes.ac.in
shettyrutar17ite@student.mes.ac.in

***Abstract* - In the current era, Digitization has taken day-to-day utilities starting from a cab to a glossary on the internet. All the service providers heavily leverage IT and IT Services. Web Application plays a significant role in providing these services. While Digital opens infinite opportunities to increase business and enhance delivery, it also exposes the business to an unseen world of cyber-attacks. To prevent the business from digital dysfunctioning, organizations pro-actively and continuously perform Vulnerability Assessments & Penetration Tests on their IT Assets (i.e. Web Applications, Network Devices, Servers, Security Devices, etc.). We propose a framework that captures the footprint of an organization, useful for the information gathering phase during penetration testing called Reconnaissance. Reconnaissance refers to the preparatory phase where a penetration tester seeks to gather as much information as possible about a target of evaluation before launching a penetration test. Our Python tool helps in locating and saving organization-specific data. Such data repositories will help in the vulnerability assessment of an organization. This will include designing a user-friendly graphical user interface. In the end, it will generate a report of vulnerability assessment.**

Keywords -- Vulnerability, Penetration, Webapp, Assessment, Reconnaissance, Footprinting.

I. Introduction

A. Fundamentals

Web Application plays a significant role in providing IT services. While the digital world opens infinite opportunities to increase business and enhance delivery, it also exposes the business to an unseen world of cyber-attacks. To prevent the business from digital dysfunctioning, organizations pro-actively and continuously perform Vulnerability Assessments & Penetration Tests on their IT Assets such as Web Applications, Network Devices, Servers. We propose a framework that captures the footprint of an organization, useful for the information gathering phase during penetration testing called Reconnaissance. Reconnaissance refers to the preparatory phase where a penetration tester seeks to gather as much information as possible about a target of evaluation before launching a penetration test.

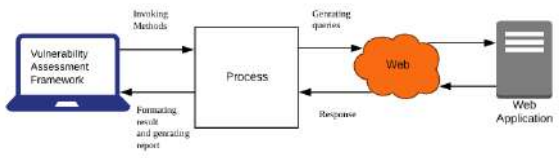


Fig 1. Vulnerability assessment framework introduction

B. Objectives

When a web application is designed and deployed it might be having vulnerabilities in it. Testing or more specifically White box testing is done before deploying an application in the production environment. While doing this job various tools and frameworks are to be used and corresponding reports are to be generated. The main objective is to build a tool that will be easy to use, integrating various other open-source tools and automating them, and generating the end report of the assessment.The main objectives to be achieved:

- To understand top 10 vulnerabilities mentioned on OWASP and how to exploit them.
- To understand frameworks like Octave.
- Developing python modules to carry out tasks of assessment.
- Implementing UI for above components.
- Test tool on DVWA and get feedback.
- Iterate and improve according to feedback.
- Generating report with help of modules and queries to DVWA.

C. Scope

Our project goal is to find and give a detailed report of vulnerabilities in the web application of IT organizations/ companies. Our project requirements are DVWA for testing, web browsers like Firefox, system installed Linux OS, Python environment, database to store reports of previous vulnerability assessment tests. Our project deliverables are Python modules of tools, UI, report module. List of users using our framework would be Penetration Testers, Cyber Security Researchers,etc. Our framework features include user-friendly UI, automating assessment tasks, detailed report generation,etc.

II. Literature Survey

In this survey, the relevant techniques in the literature are reviewed. It describes various techniques currently being used. We have reviewed four research papers in the domain of vulnerability and network security.

A. Literature review techniques

The techniques in this category are adapted to the individual needs, interests and preferences of the user or society. They are tools for suggesting items to users in this domain. Various techniques in this category are listed here. These techniques have various advantages and are used extensively in the literature.

B. Technique One

Administrators need to perform vulnerability scans periodically which helps them to uncover shortcomings of network security that can lead to devices or information being compromised or destroyed by exploits. Different tools have different approaches and outputs integrating the output and making it easy to understand. In this we are considering NMAP & OpenVAS. On the basis of impediments of NMAP and OpenVAS, another tool is developed which holds the best of both devices alongside overcoming a few drawbacks. Further vulnerabilities scanning is performed by

comparing the information obtained from a network scan to a database of vulnerability signatures to produce a list of vulnerabilities that are presumably present in the network. Along with performing network scanning and vulnerability assessment, an auto-scan mechanism is also added in a new tool to test devices when they are compromised. In other words, network mapping, vulnerabilities and configuration faults in the network are shown in various formats.

C. Technique Two

This project provides flexibility because of modular code. Tool is developed by dividing it into modules hence this makes the tool open to future development. This tool is developed to test against the web application with HTTPS hence with SSL certification only. In this technique penetration testing is also done. Net Nirikshak 1.0, Samurai framework, Safe3 scanner, Websecurify and SQLmap are automated using Python. Heterogeneous output of these tools is integrated and a report is generated. Manual testing of the vulnerabilities of the application was successfully performed. Conversion of the local server from HTTP to HTTPS was successfully done by creating a self-signed certificate. An automated tool for the vulnerability assessment of HTTPS web applications was successfully developed.The tool is currently capable of performing: Whois Scan, Basic Port Scanning, Certificate Verification, SSL Connection with the Server, Grabbing of HTTP/HTTPS links, SQLI Vulnerability Detection. The tool has been well automated, hence it does not demand any special expertise from its users, unlike other tools.

D. Technique Three

Complexity of systems and the number of systems are increasing every day. This leads to more and more vulnerabilities in Systems. Attackers use these vulnerabilities to exploit the victim's system. In this paper we proved Vulnerability Assessment and Penetration Testing as a Cyber defence technology, how we can provide active cyber defence using Vulnerability Assessment and Penetration Testing. We described the complete life cycle of Vulnerability Assessment and Penetration Testing on systems or networks and proactive action taken to resolve that vulnerability and stop possible attacks. From this paper we understand prevalent Vulnerability assessment techniques and some famous premium/open source VAPT tools. We have described the complete process of how to use Vulnerability Assessment and Penetration Testing as a powerful Cyber Defence Technology.

TABLE I. LITERATURE SUMMARY

Sr no.	Summary of literature survey		
	Paper	***Advantages***	***Disadvantages***
1	Network Scanning & Vulnerability Assessment with Report Generation by Nikita Y Jhala	Integrated two tools and output is made easy to understand	Tools still uses original modules on surface code
2	An Automated tool for Vulnerability Assessment of HTTPS Web Applications by Anand Ramesh	It is developed in such a way that it is open to further development and new functionalities can be easily added in the form of modules	This tool doesn't work with applications that does not have SSL certifications
3	Vulnerability Assessment & Penetration Testing as a Cyber Defence Technology by Jai Narayan Goel, BM Mehtreb	The survey shows combined techniques for improved performance.	It improves the user preferences for suggesting items to users.

III. PROPOSED SYSTEM

A. Overview

In today's cybersecurity world, for doing vulnerability assessment different methodologies and tools are available. These tools have specific applications and help in exploring a particular scope. There are tools to map networks, Identify underlying system architecture, visualize different nodes in business architecture, and so on. When a system is forked with this tool output is generated and this output is to be analyzed by a security expert and the end report is generated. This may lead to confusion as different tools are used to work around this process and output is heterogeneous. We are proposing a system or rather a framework to solve these issues. In this proposed framework we are going to automate various tasks in vulnerability assessment using python by integrating these existing scanning tools and provide combined classified results which will be easy to understand and hence report generation will be simplified. A unique and easy to understand interface will be provided to interact with which will make our system easy to use.

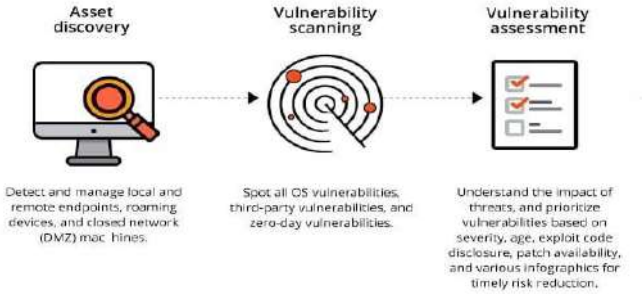


Fig 2. Vulnerability Assessment Framework proposed system

B. Existing System Tools

- OpenVAS: This is an open source tool serving as a central service that provides vulnerability assessment tools for both vulnerability scanning and vulnerability management. OpenVAS supports different operating systems.The scan engine of OpenVAS is constantly updated with the Network Vulnerability Tests OpenVAS scanner is a complete vulnerability assessment tool identifying issues related to security in the servers and other devices of the network OpenVAS services are free of cost and are usually licensed under GNU General Public License (GPL).
- Nikto: Nikto is a greatly admired and open source web scanner employed for assessing the probable issues and vulnerabilities. It is also used for verifying whether the server versions are outdated, and also checks for any particular problem that affects the functioning of the server. Nikto is used to perform a variety of tests on web servers in order to scan different items like a few hazardous files or

programs. It is not considered as a quiet tool and is used to test a web server in the least possible time. It is used for scanning different protocols like HTTPS, HTTPd, HTTP etc. This tool allows scanning multiple ports of a specific server.

- Nmap: Nmap is a handy addition to the value-added reseller (VAR) and consultants' vulnerability assessment toolbox. Nmap performs a SYN Scan, which works against any compliant TCP stack, rather than depending on idiosyncrasies of specific platforms. It can be used to quickly scan thousands of ports, and it allows clear, reliable differentiation between ports in open, closed and filtered states. If Nmap is compiled with OpenSSL support, it can even connect to an SSL server to deduce the service listening behind that encryption layer. Another advantage of running version detection is that Nmap will try to get a response from TCP and UDP ports that a simple port scan can't determine are open or filtered, and Nmap will change the state to open if it succeeds.

TABLE II. Summary of Existing System Tools

Category	Existing System Tools	
	Tool	***Description***
Host-Based	Metasploit	An open-source platform for developing, testing and exploiting code.
Network-Based	Cisco Secure Scanner	It is developed in such a way that it is open to further development and new functionalities can be easily added in the form of modules
	Wireshark	Open Source Network Protocol Analyzer for Linux and Windows.
	Nmap	Free Open Source utility for security auditing.
	Nessus	Agentless auditing, Reporting and patch management integration.
Database-Based	SQL lite	Dictionary Attack tool door for SQL server.
	Secure Auditor	Enable users to perform enumeration, scanning, auditing, and penetration testing and forensic on OS.
	DB-scan	Detection of Trojan of a database, detecting hidden Trojan by baseline scanning.

C. Proposed System

As per the previously mentioned documents, there are tons of tools or micro tools available to carry out every task. These tools are written in different languages and produce heterogeneous output. This creates confusion while analyzing these outputs and generating reports.We propose a method to solve these issues in our method. There are top 10 vulnerability lists available on OWASP also information to exploit them is available and well documented. We are going to follow this and implement it in our methodology. There are many frameworks available. These frameworks define the way to approach vulnerability assessment and penetration testing. We are going to follow one of this framework named Octave to understand how these assessments are carried out in the real world and design our workflow accordingly. This stage is very important to maintain standards as it will help security experts without causing any issues while performing tests on the system. We are going to perform the reconnaissance stage as it is considered as one of the important steps in any pen-testing session there are tools available like NMAP, Wireshark, and different APIs are provided by different developers. We are going to use these tools and available modules developed for these tools and scrap output generated. This output is then formatted according to needs and presented to the user. To scan various vulnerabilities in a web application Nikto and OpenVAS are some top-rated tools available in open source. Modules of this tool are then used to scrap output and a separate module is developed to integrate it with our framework.

D. Implementation Details

Implementation of our project is done with the help of the Python language. As python has a huge open source community and a bunch of modules to work around we have chosen Python for programming this tool. We are having huge dependencies from other software and different languages. The implementation approach will be discussed below.

E. Implementation Details

Implementation of our project is done with the help of the Python language. As python has a huge open source community and a bunch of modules to work around we have chosen Python for programming this tool. We are having huge dependencies from other software and different languages. The implementation approach will be discussed below.

F. Understanding dependency tree

The dependency tree is important as it is always needed to be satisfied while development and installing our tool. Hence we are going to maintain dependencies for our tool from different tools and as per our programming needs

G. Developing modules

We are following the modular approach in our project. This will give us freedom over the development process as each module will be fully functional and deployable parts of code. This also makes the project upgradable in the future. These modules will be based on the existing tool and modules available in the Python environment. The integration of these modules needs to be done in this stage of development which will help to call the required module when needed.

H. Developing GUI

GUI is an important part of our framework. Even though many tools available up till now are more command-line oriented and many security experts are familiar to use such tools we are considering using GUI for interactions. As a GUI it is easy to understand and navigate. We are going to use a framework for this purpose like QT Creator or Flask. This will allow the rapid development of GUI. Once the user interacts with the framework previously designed modules in python will be called or invoked.

I. Designing report generator

Report generation is the last but important phase of vulnerability assessment. In this detected vulnerabilities are ranked, flagged, and classified this will help security person or pentester to consult proper authority about the generated

report or this report can be imported and be part of the bigger picture.

IV. Block Diagram

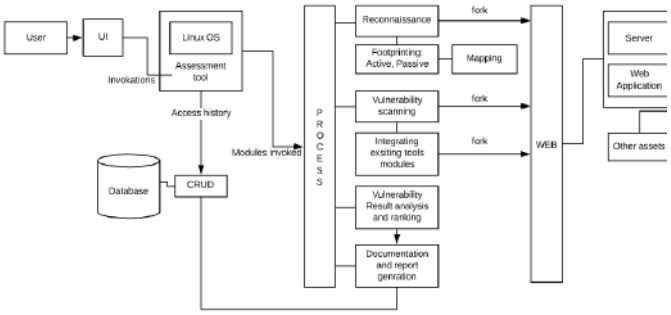


Fig 3. Block diagram of Tool

- Assessment tool: This tool will consist of the following components: different process modules, database and user interface. This tool will invoke different processes.
- Database: To store reports of previously generated reports and analysis. Thisdata can be accessed, stored, deleted using CRUD.
- Process block: This block will invoke and maintain different module calls. As a modular approach is used, the tool will be open for further development.
- Reconnaissance: This stage is for generating recon about web applications like mapping different assets available for application, discovering hidden nodes. This is done with help of footprinting modules available and designed by tool.
- Footprinting: Footprinting stage is important to map organisation resources to later iterate over them two methods will be used mainly as passive and active footprinting.
- Vulnerability scanning: In this stage vulnerabilities will be detected by using various different tools by integrated modules . Later detected vulnerabilities will be organised and ranked.
- Documentation: In this stage detected and ranked vulnerabilities will be documented according to risk level and this generated report is referred by security researchers and hence stored and retrieved from the database using CRUD.

V. 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 III and Table IV respectively.

TABLE III. Hardware Details

Processor	2 GHz Intel
HDD	180 GB
RAM	2 GB
NIC	Any

TABLE IV. Software Details

Operating System	Linux OS with GUI
Programming Language	Python, SQL
Database	MySQL
Code Editor	Visual studios

VI. Use Case Diagram

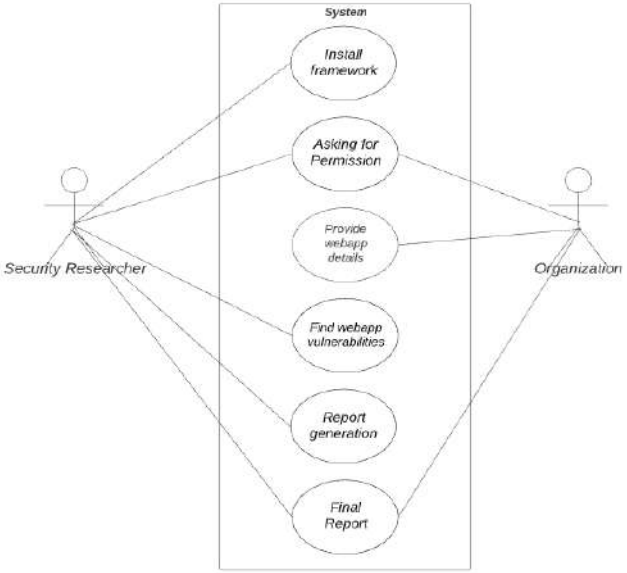


Fig 4. Use Case Diagram

VII. Summary

In this paper, we tried to explain the methodology we are going to apply while developing our framework. We analyzed what are currently tools available to a pentester and found out the advantages and disadvantages of using them standalone and how they generate output. We understood from our literature survey that how these tools can be combined together and how the output of these tools can be integrated and served to the user as one. Then we analyzed the pros and cons to put this method in action and proposed our own method. We discussed how this tool will work by looking at UML diagrams. Implementation details of this framework are discussed in brief as it can be considered as our road map to develop this tool.

VIII. Future Scope

This application framework is following a modular approach from start to finish hence we are open to changes also adding new functionality to the project. At any level, we can introduce new modules. We are planning to add more footprinting methods to applications to assess web applications better. GUI can be improved further so it can be used by a beginner and still he/she will carry out the task.

Acknowledgment

We would like to show our gratitude and thanks to Dr Sandeep Joshi, Principal, Pillai College of Engineering for giving us a good guideline for the project throughout numerous consultations. The help and guidance given by him from time to time gave us the motivation to complete the project. We also take this opportunity to express a deep sense of gratitude to Dr Satishkumar Verma, HOD, IT department, for his cordial support, valuable information and guidance, which helped in completing this task through various stages. We are grateful for the cooperation during the completion of our task. We take this opportunity to express our profound gratitude and deep regards to our guide Prof. Aju Palleri for introducing us to the methodology of work and her exemplary guidance, monitoring and constant encouragement throughout the course of this thesis. We would like to thank all the people for their help indirectly and directly to complete our task.

References

[1] Nikita Jhala (Nirma University), “Network Scanning & Vulnerability Assessment with Report Generation”.

[2] Octave framework.

[3] OWASP Top Ten vulnerability.

[4] Top vulnerability scanning tools.

[5] Nmap network scanning manual.

[6] Anand Ramesh (Institute for Development and Research in Banking Technology), “AN AUTOMATED TOOL FOR VULNERABILITY ASSESSMENT OF HTTPS WEB APPLICATIONS”.

[7] Front end development Kivy framework.

[8] Django documentation.

[9] Jai Narayan Goel, BM Mehtreb (University of Hyderabad), “Vulnerability Assessment & Penetration Testing as a Cyber Defence Technology”.

FEATURE EXTRACTION FOR GENDER AND EMOTION RECOGNITION:A SURVEY

Pooja Pillai 1, Anupama Subramanian 2, Sarah Khalife 3,Vani Nair 4, and Dr. Madhu Nashipudimath

Department of Computer Engineering, PCE, Navi Mumbai, India - 410206

Abstract— Voice recognition plays a key role in spoken communication . Humans can identify gender and emotion easily by the speech of the speaker but it is not easy for a computer.This leads to huge scope to work on speech recognition. But going through feature analysis for speech recognition is a tedious and complex task. There are many solutions to overcome this problem. Feature extraction using Mel-Frequency Cepstral Coefficient(MFCC), feature reduction by Principal Component Analysis(PCA) and Support Vector Machine(SVM) classifier to identify the gender and emotion are few solutions. Hence the motto is to find an improved approach for voice feature extraction. As a base idea, here few feature extraction techniques are discussed based on data selection, preprocessing of voice signals, feature selection and classification.

Keywords—***Voice feature extraction, data selection, preprocessing , feature selection and classification.***

1. Introduction

The rapid growth of technology and increasing human demand has made voice recognition systems one of the most desired software programs in various devices. Speech recognition technology converts spoken audio into text and lets users control digital devices by speaking instead of using conventional tools such as keystrokes, buttons, keyboards etc. Some examples of such softwares are Google voice, digital assistants, car bluetooth etc. Speech signals contain large amounts of information. Two such pieces of information are gender and emotion which can be distinguished relatively more easily by humans than by computers. At present, the research on voice recognition mainly focuses on the identification of single information, which is not enough to understand the true meaning of speech. Here we intend to use voice feature extraction to identify the gender and emotion of the person using SVM classifier, PCA and MFCC.

2. Literature Survey

A. Data Selection: The training of the proposed system will be done with the help of predefined datasets. The testing and validation of the proposed model will be done either by using predefined dataset or by taking live input from the user. The RAVDESS dataset consisting 4904 files of emotional speech in eight basic emotional categories i.e. angry, disgust, fearful, happy, calm, sad, and neutral is used [2].

B. Preprocessing voice signals: The speech signals obtained from the predefined datasets or from the user cannot be given as input directly to the feature extraction module. The input signals need to be pre-processed using Voice Activity Detection (VAD). Wang et al. [1] used this technique to select active frames and filter out silence frames which did not contain any emotional information. VAD is also used to recognize age/gender from speech. The advantage of using VAD is that even if there is long silence in the beginning or in the end of an utterance, the behavior of the classifier would not be negatively influenced.

C. Feature Extraction: For the purpose of gender and emotion recognition from speech signals it is important to extract relevant features. Feature extraction transforms the processed speech signal to a concise but logical representation that is more distinct and reliable than the actual signal.

The short term power spectrum of sound is described by Mel-frequency cepstrum (MFC), on the basis of a linear cosine transform to log power spectrum with a non-linear Mel scale of frequency. MFCC takes into account human perception for sensitivity at appropriate frequencies by converting the conventional frequency to Mel Scale. Gumelar et al.[2] used this method to perform feature extraction since it is very easy to implement and hence has become a widely used method for speech recognition. The accuracy of the system decreases if the sound samples used have low emotional intensity. It is observed that accuracy value can be increased when more datasets are involved.

Linear Prediction Coding (LPC) approximates speech samples as a linear combination of past samples. Then, by minimizing the sum of the squared differences between the actual speech samples and the linearly predicted samples over a finite interval, a unique set of predictor coefficients can be determined. Paulraj, M. P., et al.[5] presented an automatic vowel classification system based on LPC and neural network. Where traditional linear

prediction suffers aliased autocorrelation coefficients LPC gives a very accurate estimate of speech parameters and is comparatively efficient for computation. At the same time, the performance of LPC degrades on the presence of noise in audio signals.

Linear Predictive Cepstral Coefficients (LPCC) gives a stable representation of the input speech signal in compressed form as compared to LPC. LPCC are derived from the fourier transform of the log magnitude spectrum of LPC. It analyses the input signal by estimating the enhanced frequency bands by removing their effects from the signal and estimates the intensity and frequency of the remaining signal[13].

With the help of Discrete Wavelet Transform (DWT), time domain and frequency domain information of the signal can be fetched. DWT decreases the quantity of signals required to recognize the emotions. Koduru et al.[3] used different feature extraction techniques like MFCC, pitch, energy, Zero Crossing Rate (ZCR) and DWT in order to extract maximum information of the speech signal and get better accuracy. It improves the speech emotion recognition rate with less processing time.

D. Feature Selection: There are many features in a speech signal but not all of them are needed for the proposed system. Feature selection is required to extract features from audio signals for selection of principal components, as well as to remove redundancy and unused information. Principal Component Analysis (PCA) is used to find the principal components out of all available features[7]. PCA is a statistical tool which is used to convert a set of observations of correlated variables to a set of values of linearly uncorrelated variables[14]. It also reduces the processing time since a large set of information requires more processing time.

E. Classification

Emotion and gender recognition is a supervised learning problem. Each pattern used for the training of the classifier carries the correct emotion/gender class label. There is a large number of classifiers for supervised learning. The most popular approaches are Bayesian learning, the linear discriminant analysis (LDA), the support vector machine (SVM) as an extension of LDA with a high-dimensional feature space, the multi-layer neural network (NN), and the hidden Markov model (HMM) to capture temporal state transitions. SVM is the most widely used classifier due to its efficiency in classifying high dimensional data where the number of features is greater than number of observations. A significant advantage of SVMs is that while ANNs can suffer from multiple local minima, the solution to an SVM is global and unique. SVM has a simple geometric interpretation and gives a sparse solution.

2.1 Summary of Related Work

A literature review is done with respect to the associated system. The summary is given in Table 1

Table 1 Summary of literature survey

Literature	Advantages and Constraints
Sharma, Gyanendra, and Shuchi Mala [7]	Advantages: Higher accuracy is achieved by using a hybrid model consisting of PCA and SVM classifier. Constraints: Small dataset is used for training and testing the model.
Aggarwal, Gaurav, and Rekha Vig[8]	Advantages: With reduced speech features, recognition rate increased compared to existing systems. SVM achieved more accuracy than Naive Bayes. Constraints: Less no. of features were considered for classification.
Jiang, Wei, et al [11].	Advantages: Refined and unified features are fed into the fusion network module for recognition. Constraints: SVM classifier was not capable of detecting Neutral emotions correctly from audio signals.
Hossain et al. [12]	Advantages: F-1 Cepstral Coefficient method provides better performance. Constraints: Experimental dataset is not large enough and noise in audio signals is not considered here.
Pahwa, Anjali, and Gaurav Aggarwal[10]	Advantages: The proposed system removes long silence/pauses and unwanted noise for greater precision. Constraints: It considers only stored signal files for input and does not take real time input.
S. Sengupta, G. Yasmin and A. Ghosal[14]	Advantages: Functional and perceptual characteristics are used for classification. Quadratic kernel improved accuracy of SVM by 4%. Constraints: k-NN did not perform well since the value of k chosen is undecidable and there is no way to decide which value would provide best results.

Ka, Mr Sundar, et al[15]	Advantages: High computational processing and independent of ethnicity. Constraints:Accuracy is low when compared to newer techniques. It also requires some computational devices. Implementation cost is high.
Gumelar, Agustinus Bimo, et al.[2]	Advantages: The training process is repeated multiple times which increases the overall accuracy. Constraints: The results of this study will have a much more decreased accuracy if the sound samples used have low emotional intensity.
Kerkeni, Leila, et al.[16]	Advantages: 90.05% recognition rate was achieved by integrating MFCC and Modulation Spectral Features(MSF). Constraints: The RNN model has too many parameters(155 coefficients in total) and poor training data.
Manas Jain et al.[13]	Advantages: The MFCC extraction features have achieved higher accuracy compared to that of LPCC. Constraints:Accuracy of UGA dataset was low as compared to LDC dataset.

3. Existing architecture

In the existing systems, various acoustic features are extracted with the help of MFCC, LPCC, etc. The values are stored as CSV (comma-separated values) files which store tabular data in plain text and every line of the file is considered as a data record. Various models such as Random Forest, CART model, Neural networks,etc. are trained to categorise the input speech sample. For testing and validation, either predefined datasets are used or live input is taken from the user. With the help of knowledge gained after training and integrated algorithms, the gender and emotion of the speaker is classified and given as output. It is observed that deep learning neural networks or convoluted neural networks have high accuracy but are resource hungry, requiring large dataset for training and high processing time and the accuracy obtained with the help of other classification models is much less as compared to SVM classifier.

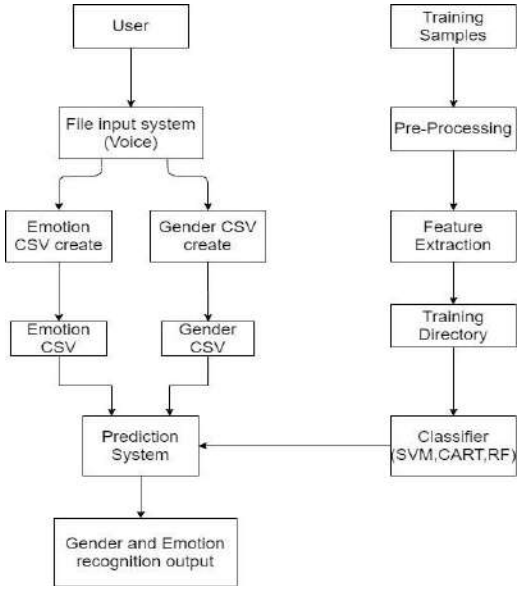


Fig 3.1: Existing system architecture

4. Proposed model

Detecting a user's emotion and gender accurately, from his/her voice input, requires complex algorithms and intricate deep learning models. To overcome this, we pre-processed the input data meticulously, followed by classification with the help of SVM, which resulted in precise identification of emotion and gender without the need of complicated Neural networks.

4.1 System Architecture

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

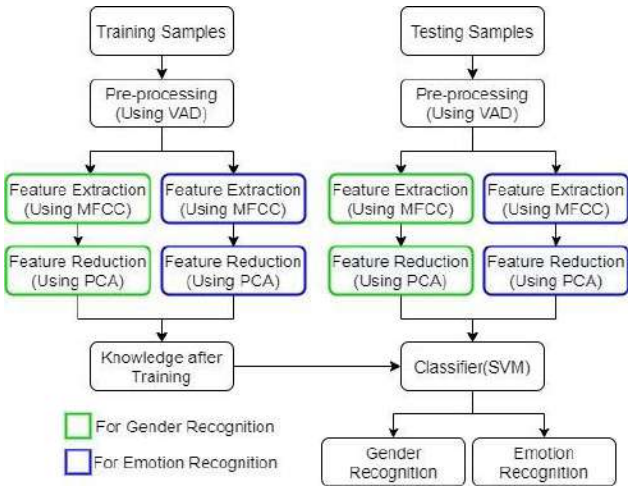


Fig. 1 Proposed system architecture

The proposed model follows the steps such as data extraction(database and real time), preprocessing signals using Voice Activity Detector(VAD). VAD helps in determining whether a particular signal contains speech or not. Mel-frequency cepstral coefficient (MFCC) can be used for feature extraction which determines unique coefficients to a particular sample after processing. Principal Component Analysis (PCA) extracts features from audio signals for selection of principal components. A binary and multiclass SVM classifier will be efficient in classifying gender and emotion from the speech signals. The proposed combination of techniques are expected to produce better results as per survey.

5. Dataset and Parameters

The sample dataset that would be used in the experiment is RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) Dataset. The database contains a total of 7356 files (total size: 24.8 GB) which consists of voice samples of 24 professional actors (12 female, 12 male) vocalizing 2 lexically-matched statements. These are in a neutral North American accent. Speech consists of happy, sad, angry, surprise, calm, disgust and fearful expressions. With additional neutral expression ,every expression is even produced at 2 levels of emotional intensity- normal and strong.

Conclusion

The study of voice feature extraction for gender and emotion recognition is presented in the aforementioned sections and previous works related to the same are studied thoroughly and analyzed for further improvements. The existing architecture is explained along with some shortcomings. The proposed model will be capable of recognizing the gender and current emotional state of a person through his/her voice input in real time. It focuses working more on the pre-processing of the input, unlike the complex Neural Network and SVM approaches in existing systems, which gives more assurance of the classification accuracy. Both Gender Recognition (GR) and Emotion recognition (ER) can be implemented using Support Vector Machine (SVM), fed with relevant audio features. Since SVM is the core in this model, it is capable of training faster with high accuracy.

ACKNOWLEDGEMENT

It is our privilege to express our sincerest regards to our supervisor Dr. Madhu M. 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 of the Department Dr. Sharvari Govilkar and our Principal Dr. Sandeep M. Joshi for encouraging and allowing us to present this work.

REFERENCES

1. *Wang, Zhong-Qiu, and Ivan Tashev.* "Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks." IEEE, 2017.
2. *Gumelar, Agustinus Bimo, et al.* "Human Voice Emotion Identification Using Prosodic and Spectral Feature Extraction Based on Deep Neural Networks." IEEE, 2019.
3. *Koduru Anusha, Hima Bindu Valiveti, and Anil Kumar Budati.* "Feature extraction algorithms to improve the speech emotion recognition rate." International Journal of Speech Technology 23.1 (2020): 45-55.
4. *Rong, Jia, Gang Li, and Yi-Ping Phoebe Chen.* "Acoustic feature selection for automatic emotion recognition from speech." Information processing & management 45.3 (2009): 315-328.
5. *Paulraj, M. P., et al.* "A speech recognition system for Malaysian English pronunciation using Neural Network." (2009).
6. *Rosenberg, A., and M. Sambur.* "New techniques for automatic speaker verification." IEEE Transactions on Acoustics, Speech, and Signal Processing 23.2 (1975): 169-176.
7. *Sharma, Gyanendra, and Shuchi Mala.* "Framework for gender recognition using voice." IEEE, 2020.
8. *Aggarwal, Gaurav, and Rekha Vig.* "Acoustic Methodologies for Classifying Gender and Emotions using Machine Learning Algorithms." 2019 Amity International Conference on Artificial Intelligence (AICAI).
9. *Rani, Poonam, and Ms Geeta.* "Gender and Emotion Recognition Using Voice."
10. *Pahwa, Anjali, and Gaurav Aggarwal.* "Speech feature extraction for gender recognition." International Journal of Image, Graphics and Signal Processing 8.9 (2016): 17.
11. *Jiang, Wei, et al.* "Speech emotion recognition with heterogeneous feature unification of deep neural network." Sensors 19.12 (2019): 2730.
12. *Hossain, Nazia, Rifat Jahan, and Tanjila Tabasum Tunka.* "Emotion Detection from Voice Based Classified Frame-Energy Signal Using K-Means Clustering." (2018).
13. *Manas Jain, Shruthi Narayan, Pratibha Balaji, Bharath K P, Abhijit Bhowmick, Karthik Ravichandran and Rajesh Muthu.* "Speech

Emotion Recognition using Support Vector Machine". (2020).

14. *S. Sengupta, G. Yasmin and A. Ghosal* "Classification of male and female speech using perceptual features," 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Delhi, 2017, pp. 1-7.
15. *Ka, Mr Sundar, et al.* "Emotion Recognition Using Support Vector Machine."
16. *Kerkeni, Leila, et al.* "Speech Emotion Recognition: Methods and Cases Study." ICAART (2).20