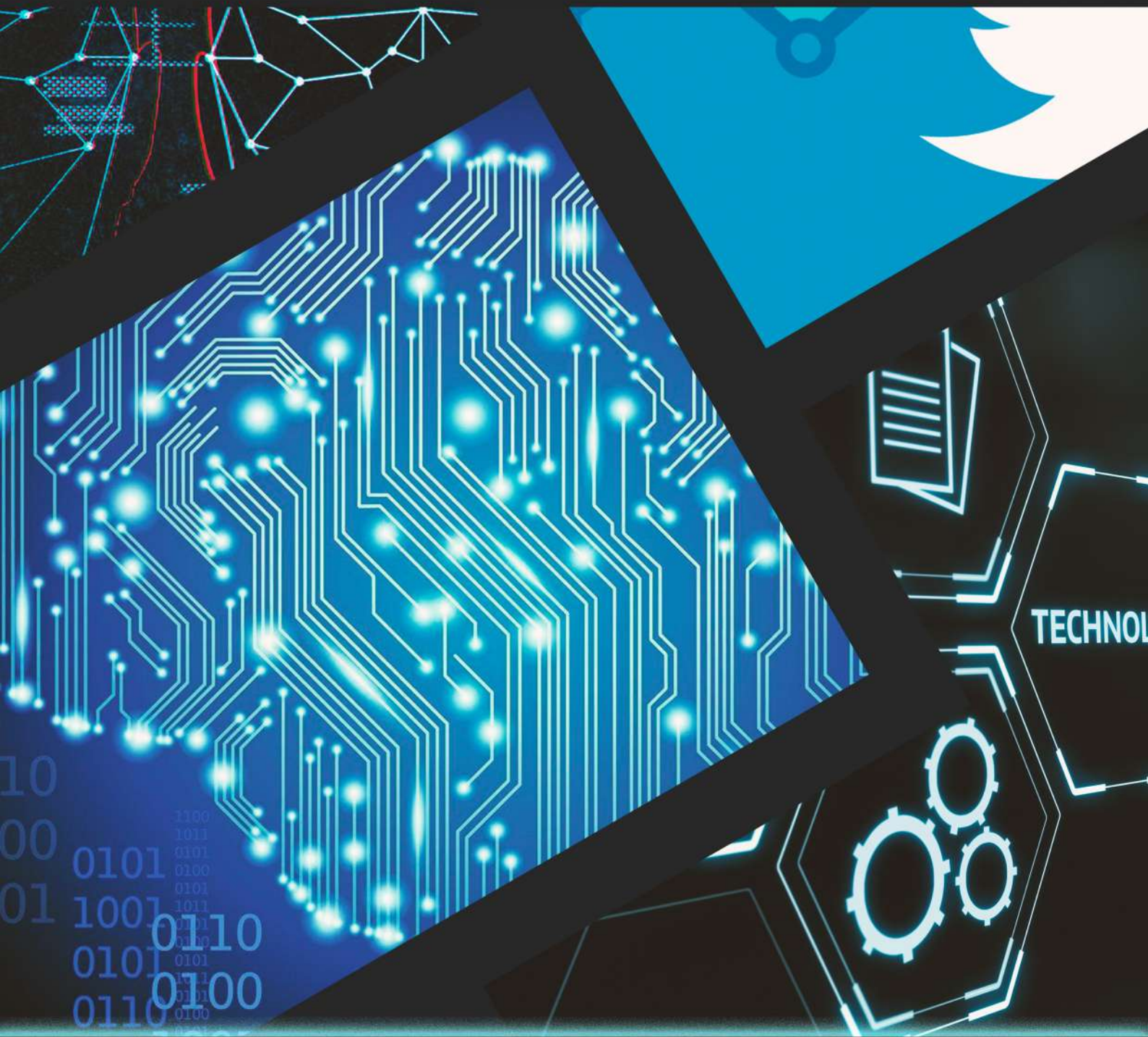




Mahatma Education Society's
Pillai College of Engineering



THE PCE JOURNAL OF COMPUTER ENGINEERING

ACADEMIC YEAR 2018-2019 | VOLUME 7, ISSUE 1

PILLAI COLLEGE OF ENGINEERING



Journal of Computer Engineering

Editors-in-Chief

Dr. Sharvari Govilkar
HOD, Dept. of Computer Engineering,
Pillai College of Engineering

Editorial Board Members

Dr. Sharvari Govilkar
Prof. Rupali Nikhare
Prof. K. S. Charumathi
Prof. Ranjita Chalke

PCE JCE _____

Volume 7

Issue 1

2018-2019

Journal of
Computer Engineering

Pillai College of Engineering

New Panvel

**Pillai College of Engineering, New Panvel.
Department of Computer Engineering
Research Papers 2018-19**

Sr. No	Title of the Paper	Page No
1	Bug Report Collection System	1-4
2	Detecting Phishing Attacks Using Natural Language Processing and Deep Learning Models	5-10
3	Automatic Fabric Defect Detection System using Image Processing	11-14
4	Text Based Emotion Analysis using Machine Learning	15-19
5	Intelligent Agriculture Greenhouse Environment Monitoring System Based on Internet of Things Technology	19-23
6	Human Posture Recognition and movement analysis Using Convolution Neural Network	24-29
7	Proposed Voice Based Notice Board Using Android	30-35
8	Secure Smart Office Automation System	34-38
9	Prediction of Movie Box-office success using NLP and Machine Learning Techniques	39-43
10	Generalized Sentiment Learner Using Deep Learning	44-49
11	Post Wi-Fi Chat Android App	50-54
12	Human Resource Analytics (HRA) For Employees Using Deep Learning	55-59
13	Thoracic Diseases Prediction Algorithm from Chest X-ray Images Using Machine Learning Techniques	60-68
14	Prediction of Indian Election Sentiments on Twitter using Machine Learning	69-74
15	Assist Crime Prevention Using Machine Learning	75-80
16	Cyberbullying Detection & Prediction in Twitter	81-87
17	Mobile Tool for analysis of Events, stocks and Management System	88-90

Editorial

It takes immense pleasure in launching this issue of the Journal of the Computer Engineering Department, PCE. The journal is a forum for the students and faculty of the department to showcase their work in various imminent fields related to computer engineering and its applications.

This issue has 17 papers comprising the outcome of research work done by the students and the faculty of the computer department, exploring the various domains such as Deep Learning, Machine Learning, Internet of Things, Natural Language Processing, Security, Image processing, Web technologies, Artificial Intelligence and others.

I hope that this issue of PCE JCE will be helpful for the future aspiring computer engineers and the research students. I thank the editorial team for their efforts put in for the launching of this issue.

Dr. Sharvari Govilkar

Editor-in –Chief

Bug Report Collection System

Aditya Mahajan¹, Kajal Dubey¹, Rahul Gupta¹, Prof. Deepti Lawand²

Department of Computer Engineering

Pillai College of Engineering

Navi Mumbai India-410206

Abstract—Bug report collection system is a "Bug tracking system" or set of scripts which maintains a database. Our system not only detects the bugs but provides complete information regarding detected bugs. The problem in the older system can be defined as the whole project maintenance, user maintenance, their assignment has to be maintained manually. The developer resolves the issues as per the requirements. In the testing phase, the tester will identify and enters the bugs into the system. Whenever the tester encounters 'n' number of bugs, description and bug priority and the bug images will be added in the database. The manager will assign bugs to the developer. The main objective of the proposed system is to fully analyze the bugs and report the same to the developer in an efficient manner so that the developer can get the right information. Bugs may be caused by tiny coding errors, but the results of bugs can be serious, finding and fixing bugs is a rather challenging task. For that purpose, our system contains a run time engine in that the code can be tested and reported to the developer. The developer will get the code snippet of that code so that the developer can easily understand the issue related to it.

Keywords: *Bugs, Bug report, software, project maintenance, tracking, Database.*

I. INTRODUCTION

This is the world of information and technologies. In this, the programmer can hardly write the programs without any bugs. It is very difficult to find the bugs and then report to the developer. For that purpose, a system is being developed to track the bugs and report to the developer. The system will be used by the testing team in the organization. There are three end users for the system that is a tester, manager and the developer. The user will register themselves by the user registration and gets the approval by the admin of the system. The purpose of the system is to collect the details of the bug and report into the system. In the existing system, the tester reports the in the form of excel document, if the document is damaged then the total information about the bug will be lost and cannot be recovered. In existing systems report generation was done manually by copying the details of different files into another file and when the manager needs the information of any bug he searches for the file into the database that was very time consuming and the information retrieval is a very big process. The bug details were not stored in the database for future reference. Due to the drawback of the existing system the project is undertaken to develop a new system providing solutions for the existing problems. Bugs may be caused by tiny

coding errors, but the results of bugs can be serious, finding and fixing bugs is a rather challenging task. For that purpose, our system contains a run time engine in that the code can be tested and reported to the developer.

II. SCOPE

Having complete information in the initial bug report (or as soon as possible) helps developers to quickly understand and resolve the bug. The focus of our work is to improve bug report collection systems with the goal of increasing the completeness of bug reports. This project not only reports identified bug but also categorize the identified bugs using priorities. The "Bug report collection system" is a web based application that can be accessed throughout the organization.

III. LITERATURE REVIEW

• Towards Effective Troubleshooting With Data Truncation:

The Effective Troubleshooting with data deals with reducing the data present in the bug repository and improve the data and reduce time and cost of bug order, it represents an automatic approach to predict a developer with enough experience to solve the new coming issue. The bug data sets are obtained and techniques such as instance selection feature selection are applied simultaneously. The top k pruning is applied for improving results of data reduction quality, obtaining domain classifier for bug triage, a necessary step is to collect numerous labeled bug reports, which are bug reports marked with their respective developers. The half supervised text approach to enhance the accuracy of bug triage. This semi-supervised approach enhances an NB classifier by n wise bug solution. Instance selection is used for finding a subset of similar instances (i.e. bug reports in bugs data). It is used to remove noise and redundant instances, Eliminate non-representative instances. Feature selection which aims to obtain a subset of relevant features (i.e., words in bugs data), Sorting of words according to feature values. [2].

• Automatic Bug Triage using Semi-Supervised Text Classification:

It generally proposes a semi-supervised text classification method for bug triage to avoid the lack of labeled bug reports in the present supervised techniques. This method generally is a mixture of naive Bayes classifier and the expectation maximization so as to take the benefit of both labeled and unlabeled bug reports. Using this the method iteratively labels numerous unlabeled bug reports and instantly trains a new set of a classifier with labels of all the different bug reports. Then it employs a weighted recommendation list to boost the performance by imposing the weights of silt achieved when using feature selection. In the paper, we have proposed a feature selection technique applicable to classification-based bug prediction. This technique is mostly used to tell the different sort of bugs in software changes, and specify the performance of the Naive Bayes and Support Vector Machine classifiers. These features include whitespace in the code added or deleted the section. This leads to a large number of features, in the thousands, and low tens of thousands. For greater project histories which generally have thousand revisions or more, this can be enhanced into hundreds of thousands of different features. The addition of many non-useful features reduces a classifier's multiple developers in training the classifier.

Before training a supervised applying expectation-maximization (EM) based on the combination of unlabeled and labeled bug reports. Initially, this method trains a classifier with labeled bug reports. After that, the approach iteratively labels the unlabeled bug reports and trains a new set of the classifier with names of all the different types of bug reports. In order to alter the bug triage, we update the approach with a weighted recommendation list (WRL) to augment the potential of unlabeled bug reports. This WRL is employed to probabilistically label an unlabeled bug report with multiple relevant developers instead of a single relevant developer.[3]

● Reducing Features to Improve Bug Prediction:

Nowadays machine learning classifiers have emerged as a way to tell the presence of a bug in a change made to a source code file. The classifier is first trained on software's previous data and then used to find the bugs. Two disadvantages of the co-existing classifier-based bug prediction are potentially deficient correctness for the practical use of a large collection of features. These giant numbers of features impact scalability and correctness of the approach. Minimizing features to Improve Bug Prediction aims in the classifier to first train on software history data, and then used to find the bugs. The disadvantage of the traditional method is that classifier-based bug predictions are generally not accurate enough for practical use, and use of a large number of features. The system uses Naive Bayes and Support Vector Machine (SVM). The system mainly gains ratio for feature selection, along with the characterization of bug prediction are accurate.

Additionally, the time required to perform classification increases with the number of features, rising to several seconds per classification for tens of thousands of features, and minutes for large project histories.[4]

3.1 Summary of Related Work

The summary of methods used in literature is given in

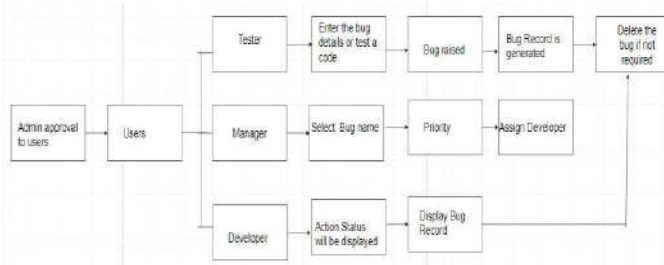
Table 2.1 Summary of literature survey

Sr no	Paper	Advantage and Disadvantage
1	Suvarnaa Kale, Ajay Kumar Gupta, "A Technique to Combine Feature Selection with Instance Selection for Effective Bug Triage", "International Journal of Science and Research (IJSR) ISSN (Online): 2319- 7064"	Advantage: It analysis data by considering the word dimension and bug dimension which helps in reducing duplicate and unnecessary bugs. Disadvantage: The order of applying instance selection and feature selection is not clearly explained which leads to inefficient system.
2	karishmaMusale, Gorakshanath Gagare, "Towards Effective Troubleshooting With Data Truncation", "International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 11, November 2015"	Advantage: The problem of handling huge number of data in bug repository is minimized. Disadvantage: Instance selection and feature selection is not completely enough to handle the data in bug repository

3	JifengXuan, He Jiang, “Automatic Bug Triage using Semi-Supervised Text Classification”, “Chinese Academy of Sciences, Beijing, 100190 China”.	<p>Advantage: It labels the bug data iteratively. The weighted list maintained, helps to boost the results obtained.</p> <p>Disadvantage: It only focuses on classifying the bugs in bug repository. The major problem in bug handling is that huge number of data in bug repository</p>
4	N. Betten burg, R. Prem raj, T. Zimmermann, and S. Kim, “Reducing Features to Improve Bug Prediction,” Proc. IEEE Conf. Software Maintenance (ICSM 08), IEEE Computer Society, Sep. 2008, pp. 337-345	<p>Advantage: Provides a layout for minimizing techniques used in bug data reduction.</p> <p>Disadvantage: Minimizing the techniques used leads in less accuracy.</p>

IV. PROPOSED WORK

A. Block Diagram:



A.Admin: The admin has the details of all the user. The admin contains the entire access of the user module. The user used to register themselves in the user registration module and get approved by the admin of the system. The admin has the right to whether to approve the user or not.

B.User: There are three types of users in the system. They are tester, manager and the developer. The user uses to register them by user registration and gets the approval by

the admin of the system. The user will be registered as per their role so they need to feel their proper details in the user registration form.

C.Tester: The bug can be a logical bug or syntax bug. For syntax bug, the tester will test the code in the run time engine(PHP code) and then report to the developer. The tester used to enter the details of the bug such as bug description,bug image in the database.Then the bug will be raised and added to the database.The tester can also delete the bug if it is not required in the system

D.Manager: The role of a manager is to assign the bug to the developer.The manager will select the bug name from the list of the bugs which was entered by the tester and the manager will assign the priority to the bugs and then assign to the respective developer.

E.Developer:The developer will get the bug details which was assigned by a manager the developer will get all the bug details such as the bug description and if the bug is logical the bug image of that particular bug will be present and if the bug is having some syntax the code snippet will be there in the report so that the developer can get the understand the issues easily where the bug is found and can proceed further for resolving the bug. The developer will give the status in the report whether the bug is the state of progress of the bug is resolved.

V. SYSTEM ARCHITECTURE

The system architecture shows the entire flow of data in our system as shown in figure 1.

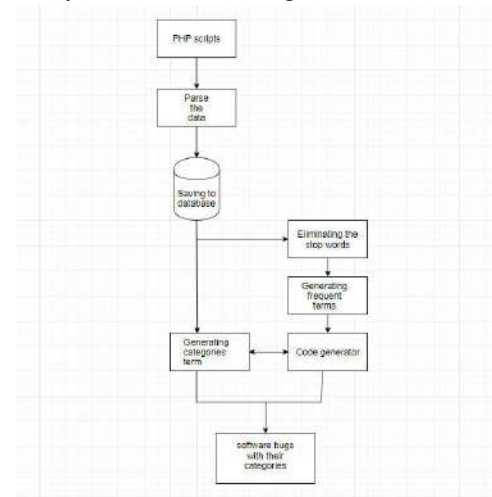


Fig. 1 Proposed system architecture

A PHP interpreter is designed to test the code and then report it to the developer so that the developer can resolve it. The PHP code will be given as input. The code will be parsed and the technique will be a top-down approach. It involves the elimination of useless words, generating frequently used terms and other involves generating categories. After all of this iteration is performed a labeled table is formed containing bug categorization on different entities. The compiler will give the output as the bug information which is extracted from the PHP code by the system. The tester will test the code and gets the bug information after testing the code. PHP is an interpreter so it will execute every line of the code and gives the output. PHP code is compiled down to an intermediate bytecode that is then interpreted by the runtime engine. After the generation of code, the bug will be categorized by frequent terms.

VI. REQUIREMENT ANALYSIS

The requirement details are given in this section.

Hardware Requirements:

- SYSTEM : Intel i5 2.8GHz
- HARD DISK : 1Tb
- RAM : 4Gb

Software Requirements:

- Operating System:
- Programming language: HTML, CSS, Javascript, PHP
- Database: PhpMyAdmin

VII. APPLICATION

● Mobile apps

Developing mobile applications is a tricky task. Even in the closed iOS market, there are many different devices and countless versions of the operating system. Making sure your app runs smoothly on everyone's device can be next to impossible without a massive testing team. Apps that crash, or don't function as they should, are certain to receive poor reviews in the App Store. Due to this bug tracking system it would be easy to track down all the bugs in an application and resolve it.

● Various industries

Sometimes we have to handle various sites at a time in a organization. There is a possibility that this server may crash. This will happen because of the bugs that are present in the system. To handle this bug tracking is necessary that will be of major help in industries

VIII. CONCLUSION

This paper helps the software concern to detect and

manage the bug in their products effectively and efficiently. With the ability to provide comprehensive reports, documentation, searching capabilities, tracking bugs and issues, bug report collection system software is a great tool for those software development needs. The main aim of this project was to identify the syntax errors and to report to developer. For that purpose a run time engine is designed in that the code will be tested and then the report will be generated. Many applications of these domains are also identified. The existing system of this project was explained which mainly was a drawback before. But now the proposed system has been a change in this domain. This proposed system has been explained in this paper with also the specifications that will be done in future with the help of science.

Acknowledgment

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

We are thankful to Dr. Sharvari Govilkar, H.O.D, Computer Engineering Department and Prof. Gaurav Sharma, B.E. Project Coordinator, Pillai college of Engineering, New Panvel, for his guidance, encouragement and support during our project. It is a great pleasure and moment of immense satisfaction for us to express our profound gratitude to our Project Guide, Prof. Deepti Lawand, whose constant encouragement enabled us to work enthusiastically. Without his encouragement this paper wouldn't have been published.

References

- [1] Suvarnaa Kale, Ajay Kumar Gupta, "A Technique to Combine Feature Selection with Instance Selection for Effective Bug Triage", "International Journal of Science and Research (IJSR) ISSN (Online): 2319- 7064", Inpress
- [2] karishmaMusale, Gorakshanath Gagare, "Towards Effective Troubleshooting With Data Truncation", "International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 11, November 2015", Inpress
- [3] JifengXuan, He Jiang, "Automatic Bug Triage using Semi-Supervised Text Classification", "Chinese Academy of Sciences, Beijing, 100190 China", Inpress
- [4] N. Betten burg, R. Prem raj, T. Zimmermann, and S. Kim, "Reducing Features to Improve Bug Prediction," Proc. IEEE Conf. Software Maintenance (ICSM 08), IEEE Computer Society, Sep. 2008, pp. 337-345, Inpress

Detecting Phishing Attacks Using Natural Language Processing and Deep Learning Models

Fenny Zalavadia, Akshata Nevrekar, Priyanka Pachpande, Shubhangi Pandey, and Dr. Sharvari Govilkar

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

Abstract— Phishing attacks are very common but least defended security threats today. An approach is proposed which uses Natural Language Processing techniques to analyze text and detect inappropriate statements which are indicative of phishing attacks. NLP offers a natural solution for this problem as it is capable of analyzing the textual content to perform intelligent recognition and performing semantic analysis of text to detect malicious intent. The approach will also use Deep Learning frameworks with hierarchical long-short term memory networks (H-LSTMs) and attention mechanisms to model the emails simultaneously at the word and sentence level. Phishing attacks categorizes the emails based on certain properties which give more details about the source of phishing. Generally, most of the existing systems focus on email classification depending upon header part or body part.

Keywords— Phishing detection, SVM, Naive Bayes, machine learning, email fraud, neural network, LSTM.

1. Introduction

Phishing takes place when cybercriminals send malicious emails designed to trick individuals into falling for a scam. The intent is usually to urge users to reveal financial data, system credentials, or alternative sensitive information. The term “Phishing” came about in mid 1990’s, when hackers began using fraudulent emails to fish for information from unsuspecting users. Cybercriminals use phishing because it’s simple, low cost and effective. Email addresses are simple to get and emails are free to send. With very little effort and small price, attackers will quickly gain access to valuable information. We can detect these emails and detect them as spam and reduce these attacks. To do this we can use various machine learning and deep learning models.

In Oct 2003, Paypal users were hit by the Mimail virus; after they clicked on a link contained in a

phishing email, a popup window pretending to be from Paypal opened and asked them to enter their user/password, that was instantly sent to the hackers. In 2004, potential voters for presidential candidate John Kerry received an official-looking email, encouraging them to donate via an enclosed link; it turned out to be a scam operative in India and Texas that had no affiliation to the Kerry campaign.

Today, strategies of phishing are as varied as, well, fish within the sea; fraudsters still come back up with new ways that to achieve trust, avoid detection, and bring disturbance. One among several troubling trends is that the use of data gleaned through social media to form the communications as personal as possible, generally cited as “spear-phishing” or “social engineering fraud.”

2. Literature Survey

1. Random Forest:

Andronicus A. Akinyelu, Aderemi O. Adewumi proposed a classifier that with better prediction accuracy and fewer numbers of features. From a dataset consisting of 2000 phishing and ham emails, a set of prominent phishing email features were extracted and used by the machine learning algorithm with a resulting classification accuracy of 99.7%[2]. Srishti Rawal, Bhuvan Rawal, Aakhila Shaheen, Shubham Malik discussed phished email classifier in which 9 features were extracted from all emails in a self-made dataset which consists of n phished emails and m ham emails[9].

2. Support Vector Machine:

Fergus Toolan, Joe Carthy [1] the instances provided by him is very small consisting of only five features. Results of an evaluation of this system, using over 8,000 emails approximately half of which were phishing emails and the remainder legitimate, are presented. Adwan Yasin, Abdel Munem Abuhasan [6] proposed a model that applied the knowledge

discovery procedures using five popular classification algorithms among which SVM was one and achieved a notable enhancement in classification accuracy. Srishti Rawal, Bhuvan Rawal, Aakhila Shaheen, Shubham Malik [9] Aim is to use the least number of features to develop a system which provides higher accuracy and study the variation of features. The features were extracted using regular expressions and NLTK. Maximum accuracy of 99.87% is achieved in classification of emails using SVM.

3. Naive Bayes:

Adwan Yasin, Abdel Munem Abuhasan [6] introduces the concept of phishing terms weighting which evaluates the weight of phishing terms in each email. The pre-processing phase is enhanced by applying stemming and WordNet ontology to enhance the model with word synonyms. Elif Yerli, Ibrahim Sogukpinar [7] discussed a technique from which success rate of 89% has been achieved against phishing attacks coming from email messages. Srishti Rawal, Bhuvan Rawal, Aakhila Shaheen, Shubham Malik [9] discussed a model where 9 features were extracted from all emails in a self-made dataset which consists of n phished emails and m ham emails. These features are given to the classifiers and results noted.

4. LSTM:

Minh Nguyen, Toan Nguyen, Thien Huu Nguyen [10] presented a framework with hierarchical long short-term memory networks (H-LSTMs) and attention mechanisms to model the emails simultaneously at the word and the sentence level. Expectation is to produce an effective model for anti-phishing and demonstrate the effectiveness of deep learning for problems in cybersecurity. The precision, recall and F1-score are used to evaluate the performance of the models for detecting phishing emails are compared with the SVM baselines in two different settings when the email headers are not considered. 2 types of data: without header and with header. Without header accuracy of 98.1% and With header accuracy of 99%.

2.1 Summary of Related Work

After going through most of the research papers from 2014 to 2018 on the topic Email Phishing detection we can infer that mostly the dataset that are used are Spamassassin and Phishing Corpus and these are widely open sourced dataset and easily available. The ML techniques mostly used till date are SVM, Random Forest, Naive Bayes, Logistic Regression, Clustering and the latest research papers have used DL techniques such as Neural networks and LSTM.

A few of the papers have encountered a high accuracy but on small set of data. In 2014 Paper Clustering technique is used, which has acquired good clusters but interpreting those cluster behavior is a bit difficult. Whereas on other hand we can observe that algorithms such as SVM, Random Forest have outperformed on various datasets and provided accuracy above 99%.

3. Proposed Work

We propose a system that takes input as standard dataset or mixture of datasets with good amount of data for example Spamassassin or phishcorpus dataset that are available as an open source data or we can create self owned dataset by collecting the emails, and apply machine learning and deep learning techniques such as LSTM, Naive Bayes, SVM, Random Forest etc and then extract features from them such as Tag, Url and many more which are fed to the models so as to classify the email as phished or legitimate. In the existing systems, a small sample dataset is taken and hence good accuracy is encountered. Firstly, using techniques such as clustering, for determining the behaviour of the resultant clusters becomes difficult.

3.1 System Architecture

The system architecture is given in Figure 1. Each block is described in this Section. The architecture consists of several stages like email dataset, Preprocessing, Feature extraction and then applying different Machine learning and Deep learning models and finally classifying the given email as legitimate or phished.

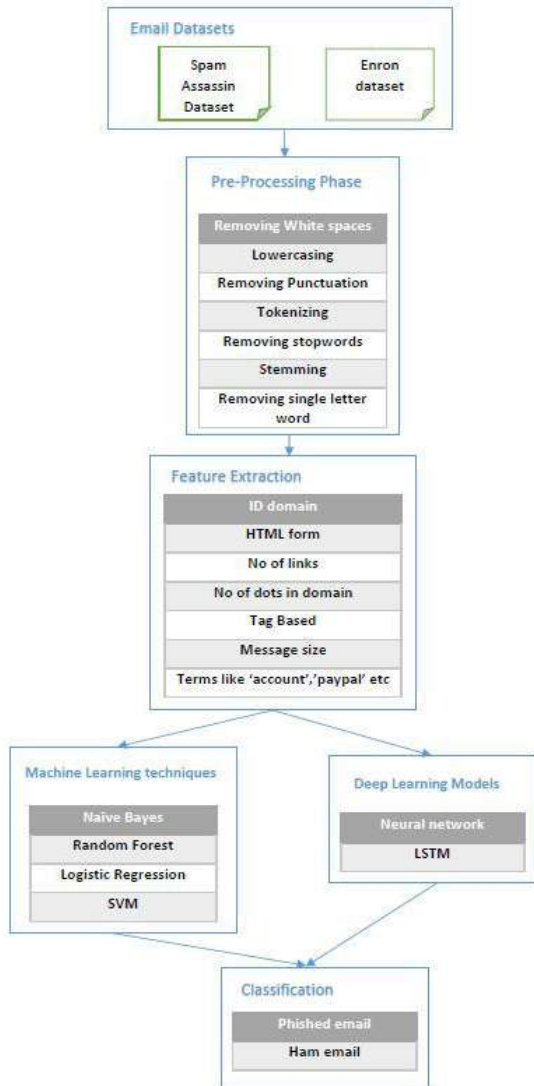


Fig. 1 Proposed system architecture

1. Email Dataset: An experiment is conducted in order to identify the input/output behavior of the system. Identify inputs. Specify the sample inputs that would be used in the experiments. The dataset collected in the experiment are identified and given in Table 1.

Dataset	Total	Phished	Legitimate
SpamAssassin	6047	1897	4150
Trec 2007	75,419	50,199	25,220
Self-Created	4,170	1,170	3,000

Table. 1 Dataset Collected for Experiment

2. Pre-Processing:

This section describes the pre-processing techniques applied on raw dataset.

2.1. Removal of Whitespace

Clean text often means tokens or a list of words that we can work with in our machine learning models. This means converting the raw text into a list of words and saving it again. A very simple way to do this would be to split the document by whitespace, including “”, newlines, tabs and few more. We can achieve this in Python with the split() function on the loaded string.

2.2. Removal of Punctuation

In the process of removing of punctuation, we first define a string of punctuation. Then we need to iterate over the provided string using a for loop. In each iteration, we check if the character is a punctuation mark or not using the membership test. We have an empty string to which we add (concatenate) the character if it is not a punctuation. Finally we display the cleaned up string.

2.3. Tokenization

Process of changing sentence into a series of words so that processing word by word can be easily performed. Given a sequence of character and a defined document unit, tokenization is the task of dividing it up into items, known as tokens, maybe at same time discarding characters, like punctuation. We tend to use white space character for tokenization.

2.4. Removal of Stopwords

Stop words are words which are not of much significance to be used in Search Queries. Most of the search engines are programmed to ignore the stop words. Table 3.2 shows sample of stop words.

Table 2 shows sample of stop words.

myself	theirs	being	a
an	the	but	for
could	let's	that's	who's
what's	at	by	with
about	against	between	into
through	during	before	after

Table 2 Sample of stop words dictionary.

2.5. Stemming

Stemming makes an attempt to get rid of the variations between inflected forms of a word, so as to scale back every word to its root form. Stemming can be performed using two approaches: the dictionary based approach and porter stemming algorithm.

3. Feature Extraction

3.1. Link based

Domain count: In order to make the links look legitimate, attackers/hackers add subdomains to these links. Adding subdomains to the links, increased the number of dots in the link.

Number of links: As compared to ham, phished emails generally contain greater number of links since the sender aims to redirect the user to an illegitimate website by deceiving him. This is a continuous feature.

3.2. Tag based

Presence of javascript: Presence of javascript in an email suggests that the sender is either trying to activate certain changes in the browser or hide information.

Presence of form tag: In order to obtain personal details from users, phished emails contains forms in them. This is a binary feature i.e. the form tag indicates that it is a phished email.

3.3. Word based

Number of action words: Presence of certain action words in emails indicates whether the attacker is expecting a response from the user to carry out certain action such as filling a form, clicking on a link, providing certain information etc. This is a continuous feature.

Presence of word account: This would suggest that the email is searching for email related to an account. It can be a bank account or social media account etc. It is a binary feature.

Presence of word paypal: Often, the attacker pretends to be a part of organization which seem legitimate. Presence of the word paypal in the “from” section or in the links of the email or would suggest that the attacker is associated with paypal. This is a binary feature.

4. Classifiers

4.1. Machine Learning Techniques

4.1.1. Naive Bayes Classifier

It is a classification technique which uses Bayes' Theorem with an assumption of independence among predictors. In simple terms, this classifier assumes the presence of a specific feature in a class is not related to the presence of other feature. Even if these features are dependent on each other or on the existence of the other features, all of these properties

independently contribute to the probability. Naive Bayes model is easy to build and useful for very large data sets.

4.1.2. Random Forest

Random forest or random decision forest are an ensemble learning method used for classification, regression and other similar tasks, that operate by constructing a decision tree at training time and output the class that is the mode of the classes that is classification or mean prediction of the individual trees.

4.1.3. Logistic Regression

It is a statistical procedure for analysing a data set in which there are one or more independent variables that determines an outcome. The result is measured with a dichotomous variable. The goal of logistic regression is to find out the best fitting model to explain the connection between the dichotomous characteristic of interest and a set of independent variables.

4.1.4. SVM (Support Vector Machine)

SVM are supervised learning models with associated learning algorithms that analyze data which is used for classification and regression analysis. Given a set of training examples, each of them marked as belonging to at least one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to at least one category or the other, making it a non-probabilistic binary linear classifier.

4.2. Deep Learning Techniques

4.2.1. Neural Network

A neural network consists of neurons, arranged in layers, which processes an input vector to give some output. Each unit takes an input, applies certain function to it and then passes the output on to the next layer. Generally the networks are defined as feed-forward: a neuron feeds its output to all the units on the next layer, but there is no feedback to the previous layer. Weightings are then applied to the signals passing from one unit to another, and these weightings are tuned in the training phase to adapt a neural network to tackle particular problem at hand.

4.2.2. LSTM

Long short-term memory (LSTM) are units of a recurrent neural network (RNN). An RNN composed of LSTM units is commonly referred as LSTM network. A standard LSTM unit consists of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the 3 gates regulate the flow of data into and out of

the cell. Long short-term memory networks are well-suited to classify, process and making predictions based on time series data, since there can be lags of unknown duration between important events in a time series.

4. Requirement Analysis

The experiment setup is carried out on a computer system which has the different software and hardware specifications as given in Table 4.1 and Table 4.2 respectively.

4.1 Software

Operating System	Windows 10
Programming Language	Python
Libraries	Pandas, Numpy, Scikit, matplotlib, keras

Table 4.1 Software details

4.2 Hardware

System	Intel core i7 generation
Processor Speed	2.2 GHz
GPU	4GB or more
RAM	8GB or more

Table 4.2 Hardware details

5. Evaluation Metrics

The different performance metrics that can be used for evaluating our model: Precision, Recall, F-measure, Confusion matrix, Accuracy.

ACKNOWLEDGMENT

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

REFERENCES

1. Fergus Toolan, Joe Carthy, "Phishing Detection using Classifier Ensembles," 2009 eCrime Researchers Summit, Tacoma, WA,

2. Andronicus A. Akinyelu, Aderemi O. Adewumi, "Classification of Phishing Email using Random Forest Machine Learning", Vol.2014, Article ID 425731, Hindawi Publishing Corporation, April 2014.
3. Sowndarya Karri, SSSN.Usha Devi N, "Framework for Phishing Detection in Email under Heave using Conceptual Similarity", Vol.2 Issue:8, International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC), August 2014.
4. Shivam Aggarwal, Vishal Kumar, S D Sudarsan, "Identification and Detection of Phishing Emails using Natural Language Processing Techniques", 2015.
5. Simranjit Kaur Tuteja, Nagaraju Bogiri, "Email Spam Filtering using BPNN Classification Algorithm", 2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT), 2016.
6. Adwan Yasin, Abdel Munem Abuhasan, "An Intelligent Classification Model for Phishing Email Detection", Vol.8, No.4, International Journal of Network Security & Its Applications (IJNSA), July 2016.
7. Elif Yerli, Ibrahim Sogukpinar, "Email Phishing Detection and Prevention by using Data Mining Techniques", 2017.
8. Naghmeh Moradpoor, Benjamin Clavie, Bill Buchanan, "Employing Machine Learning Techniques for Detection and Classification of Phishing Emails," Computing Conference 2017, London, UK, pp 149-156, July 2017.
9. Srishti Rawal, Bhuvan Rawal, Aakhila Shaheen, Shubham Malik, "Phishing Detection in Emails using Machine Learning," Vol.12 – No. 7, International Journal of Applied Information Systems (IJ AIS), October 2017. 42
10. Minh Nguyen, Toan Nguyen, Thien Huu Nguyen, "A Deep Learning Model with Hierarchical LSTMs and Supervised Attention for Anti-Phishing", May 2018.
11. Tianrui Peng, Ian G. Harris, Yuki Sawa, "Detecting Phishing Attacks using Natural

Language Processing and Machine Learning,” 12th IEEE International Conference on Semantic Computing, 2018.

AUTOMATIC FABRIC DEFECT DETECTION SYSTEM USING IMAGE PROCESSING

1.Rutuja Mapari. 2.Susmitha Nadar. 3.Uzma Naik. And Guide : Prof. Rupali Nikhare

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

Abstract

For a long time fabric defect detection is carried out manually with human visual inspection. However, they are not able to detect more than 60% of defect. Automated visual inspection systems are highly needed in the textile industry. Inspection and defect detection is important for quality control purpose. Automatic fabric detection is important for fabric analysis on the basis of digital processing to find out whether the fabric is defect free or defected. Image acquisition device is used to acquire digital fabric images. Neural networks can be used to extract patterns and detect trends that are too complex to be noticed by humans. Using neural networks as a classifier requires 2 phases, namely training phase and testing phase. The advantage is to get a warning when a certain amount of defect or imperfection occurs so that precaution measures can be taken before the product is sent to the market. (Keywords: Fabric defect; Computer vision; Defect classification; Structural approaches; Feature Extraction; Performance metrics ;Artificial Neural Network (ANN);SVM; Particle Swarm Optimization(PSO).)

1.Introduction

In textile industry, fabric defect detection plays an important role in the quality control.

Defect detection or inspection is a process identifying and locating defects. A fabric defect is a result of the manufacturing process. The textile industry is very concerned with quality. It is desirable to produce the highest quality goods in the shortest period of time possible. The quality of the fabric can be improved by decreasing defects in the fabric. Fabric Defects Fabric texture refers to the feel of the fabric. It is rough, velvety, smooth, soft, silky, lustrous, etc. The different textures of the fabric depend upon the types of weaves used. Textures are given to all types of fabrics, cotton, silk, wool, leather, and also to linen. Textile Fabric materials are used to prepare different categories and types of Fabric products in the textile industry. Natural fabric and synthetic fabric are the two different classifications of textile fabric. Synthetic fabrics are fairly new and have evolved with the continuous growth in textile industry.

2.Litarature survey

A.B.Karunamoorthy,Dr.D.Somansundareswari and S.P. Sethu[1] proposed an Artificial Neural Network(ANN) based approach. First stage was Image Preprocessing. Second stage was Image Decomposition in which cartoon was attained by executing ID on the preprocessed patterned images. Third stage was Detection Enhancement in which the carton image was converted into a binary image whose 1-valued pixels represent

defective objects while 0-valued pixels are defect free regions.

B. H Ibrahim Celik, L Canan Dulger and Mehmet Topalbekiroglu proposed fabric defect detection using linear filtering and morphological operations. The algorithm is applied off-line and real-time to denim fabric samples for five types of defects(warp,lacking,hole, soiled yarn and knot). As the fabric was wound, image frames were captured and they were analyzed on the computer. The fabric motion and image acquisition process was synchronized with a rotary encoder via a frame grabber card. Linear filters are used to segment the defective region. The filtered image is converted to binary image using Double thresholding method. Then applied to dilation process and then the remaining noises are removed via erosion. The, feature extraction is performed using Discrete wavelet transform. Finally, the defects are classified using FFNN method.

C. Srinadh Unava, Kirankumar jetti and MVSS Nagendranath proposed Fabric Fault Detection system using Regular bands and ICA. ICA was used for indicating and locating the defects on patterned fabric images. Two sets of features named statistical features and texture features was introduced. Statistical features such as mean, standard deviation , variance, coefficient of variation, skewness and kurtosis was used to characterize the histograms and distinguish between normal and defective fabrics. Texture features include a large group of shape description techniques such as regularity, elongation,direction,compactness etc.

D. Mark et.al proposed a prototype of real-time computer vision system for detecting defects in fabrics. They proposed a filter selection method which could automatically tune the Gabor functions to match with the

texture information. The defect types studied were harness, breakdown, miss pick, warp burl and water damage.

2.1 Summary related work

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

Sr. no.	Literature	Advantage
1.	H Ibrahim Celik, L Canan Dulger and Mehmet Topalbekiroglu	Hole defect is recognized with 100% accuracy rate, the others are recognized with a rate of 95%.
2.	Srinadh Unava, Kirankumar jetti and MVSS Nagendranath	ICA model is noise free. It helps in improving efficiency.
3.	B.Karunamoorthy ,Dr.D.Somansundareswari and S.P. Sethu	ANN classifier separates the faulty fabric from fault free fabrics. ANN shows a promising 95% accuracy with just 20 samples.

3.Proposed work

Automatic Defect detection system using image processing require fast and effective algorithm ,that brings us to develop algorithm which require less execution time and still get more accurate results and avoid unnecessary calculation.

3.1 System Architecture:

The system overview gives a brief description about the overall working of the system. Here, the user interacts with the system by providing a dataset of fault free and faulty images. The further processing is explained below.:

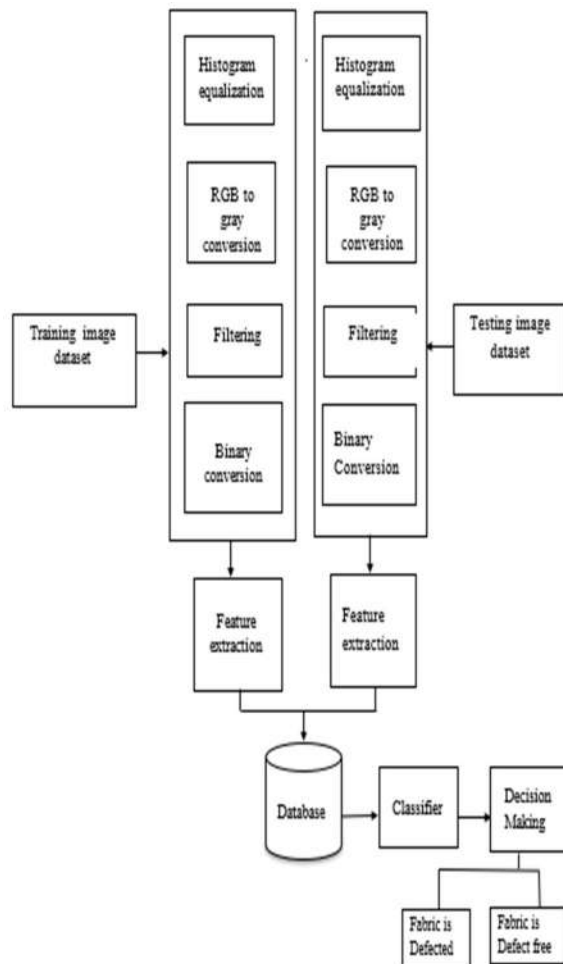


Fig.3.3 Proposed system architecture

A. Image acquisition : In this process the camera is used to take the picture from the area of interest. The acquired image is saved then and is helpful for the further process. The acquired image is then proposed to the MATLAB software in which the image is stored for the further processing.

B. Image Preprocessing : Image preprocessing simply means that the resize the image, histogram equalization and noise removing etc. In the images the noise is random alteration in the energy of an image that can be simply removed by using different filtering approaches.

The image filtering helps in various kind of applications, such as smoothing, sharpness, noise removal , and edge detection .

C. Feature Extraction : Providing the input data into the group of features is called Feature Extraction of image .Image characteristic gives the useful information about an image and rejects the rest. Feature Extraction is a stage in which various methods can be employed for capturing visual content of images for indexing & retrieval purpose. There can be number of features defined from an image and there are methods for calculating each of these features. The features which are better suited for a particular application are selected for further analysis .

D. Classifier : Image classification is most important part of image analysis. Classification is nothing but group the similar types of object and dissimilar type of object into a different partition, with the aim to providing a easy way for image analysis. Artificial Neural Network(ANN), support vector machines (SVMs), clustering and statistical inference are to name some effective classifiers. The classification stage gives the end result of the entire fabric defect detection process by reporting whether the fabric is defected or defect free. Using neural networks as a classifier requires two phases namely, a training phase and a testing phase. In the training phase, the neural network

makes the proper adjustment for its weights (W) to produce the desired results.

E. Making Decision : The resultant output which tells us whether the fabric which is examined is contain any kind of defect or else it is defect-free.

3.2.5 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 3.1 and Table 3.2 respectively.

Table 3.1 Hardware details

Processor	2 GHz Intel
HDD	180 GB
RAM	2 GB

Table 3.2 Software details

Operating System	Windows XP Professional With Service pack 2
Programming Language	Python
Database	Oracle 9

Acknowledgement

We would like to take this opportunity to express our gratitude sincere thanks to Prof. Rupali Nikhare for providing guidance that

she gave owing to her experience in this field for past many years. She had indeed been a lighthouse for us in this journey. We extend our sincere appreciation to all our professors from our college for their valuable insights and tips during the designing of the project.

References

[1]Rao R K Ananthavaram, O.Srinivasa Rao and MHM Krishna Prasad, “Automatic Defect Detection of Patterned Fabric by using RB Method and Independent Component Analysis”, International Journal of Computer Applications 39(18):52-56, February, 2012.

[2] Farida S.Nadaf, Nayana P.Kamble, Rohini B.Gadekar, “Fabric Fault Detection Using Digital Image Processing”, International Journal on Recent and Innovation Trends in Computing and Communication ,ISSN: 2321-8169 Volume: 5 Issue: 2 128 – 130, February, 2017.

[3] S. Sahaya Tamil Selvi1, G. M. Nasira2, “An Effective Automatic Fabric Defect Detection System using Digital Image Processing”, J. Environ. Nanotechnol Volume 6, No.1(2017) pp. 79-85 ISSN, doi:10.13074, 2017.

[4] G M Nasira,P Banumathi, “Fourier Transform and Image Processing in Automated Fabric Defect Inspection System”, International Journal of Computational Intelligence and Informatics, Vol. 3: No. 1, ISSN: 2349 – 636361, April – June, 2013.

[5] Prof. P. Y. Kumbhar, Tejaswini Mathpati, Rohini Kamaraddi and Namrata Kshirsagar, “Textile Fabric Defects Detection and Sorting Using Image Processing” International journal for research in emerging science and technology, Volume-3,Issue-3,E-ISSN:2349-7610, Mar, 2016.

Text Based Emotion Analysis using Machine Learning

Saurabh Mete, Mandira Adak, Mayuri Chilekar, Raj Bhanvadia, and Sagar Kulkarni

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

Abstract— Nowadays, detecting emotional state of a person by analyzing a text document written by him/her appear challenging but also essential many times. Emotional analysis is the process of finding out the emotions behind a sentence which can then be used to know the state of mind of the users and understand the attitudes, opinions and emotions expressed. The method of text emotion classification based on machine learning can be used to classify the input natural language texts into different categories of emotions. The machine learning approach initially takes in a few statements as a part of testing and returns the emotions associated with it. The machine learning approach constantly updates its dataset and is much more efficient if sufficient testing is done. Our proposed system can be used to reduce the cyber bullying, can be used by police authority to investigate the suicide letter and in many other such critical situations. It can also be used for product recommendation system and analysis.

Keywords—Attitudes, Emotion, Cyber bullying

1. Introduction

The problem which people are facing nowadays the emotions behind the text are unrecognised so the emotion analysis system helps to understand the emotions which are hidden in the text.

Emotion Analysis (EA) is a task which finds orientation of one's opinion in a piece of information with respect to an entity. It deals with analysing emotions, feelings, and the attitude of a speaker or a writer from a given piece of information. Emotion Analysis involves capturing of user's behaviour, likes and dislikes of an individual from the text.

A great body of work exists in the field of emotion extraction. The work done in this area includes distinguishing subjective portions in text, finding sentiment orientation and, in few cases, determining

fine-grained distinctions in sentiment, such as emotion and appraisal types. Work exclusively on emotion detection is comparatively rare and lacks empirical evaluation.

2. Literature Survey

A. Sep-tune-eval method: Kashif Khan, Sher Hayat and Muhammad Ejaz khan extracted feature sets and then used those feature sets to run the algorithm using sep-tune-eval. He use 90% of his dataset for training and 10% for testing. He also divided the dataset into two sets.

1. Sentences with neutral emotions
2. Sentences with proper emotions

The measured accuracy which he got by using neutral and positive emotions is 63% .[4]

B. Emotion Detection based on NLP: Prof. Hardik S. Jayswal, Dhruvi D. Gosai and Himangini J. Gohil used natural language processing for the text emotion detection. They classified the input text into different emotions by finding the emotional content from the given English text. The source of input to the system was textual content from social networking websites such as product reviews, comments, personal blogs, feedbacks etc.

Then they started pre processing the text with the methods such as removing punctuation, repeated characters, unwanted words, Stemming and Lemmatization. They created a dictionary which contained the word and their respective emotions such as happy, sad, fear, anger, disgust and surprise. Then the sentimental measure is done by calculating the frequency of words having the happy, sad, fear, anger, disgust and surprise tags [1].

C. Pointwise Mutual Information parameter: Kaitlyn Mulcrone used an unsupervised machine learning

algorithm for the text emotion detection. Since YouTube comments are a rich resource of natural expressions of feelings, thoughts and opinions the researcher uses them. They used statistical measures in order to compute the semantic relatedness between words of a given sentence and the target emotions. This method does not require a pre-trained dataset. It consists of measuring the Pointwise Mutual Information parameter (PMI) between each word in the text to classify and representative words of each target emotion. This measurement is based on the co-occurrence between the word to classify and the representative words in the corpus. This system achieves an average precision of 92.75%, and 68.82% as average accuracy which is close to measures given by previous systems, using SVM as machine learning algorithms [5].

D.Sentiment Classification using Machine Learning:

The sentiment about movie review is discussed in this paper. The data taken was divided into two sub datasets which consisted 700 positive and 700 negative reviews. The individual review data was referred as “unigram” and the combination of the sub datasets where referred as “bigram”. Then three standard algorithms: Naive Bayes classification, maximum entropy classification, and support vector machines were used. Below table shows the results obtained.[3] We can see that SVM attained the most accurate results.

E.Emotion prediction using NLP

The text-based emotion prediction problem in the domain of children’s fairy tales using NLP, with child-directed expressive text-to-speech synthesis as goal.The researcher extracted feature sets and then used those feature sets to run the algorithm of sep-tune-eval. He use 90% of his dataset for training and 10% for testing. He also divided the dataset into two sets.

1. Sentences with neutral emotions
2. Sentences with proper emotions

[6]After he implemented his method he got the accuracy as 63% for both neutral and proper emotions.

F.Emotion Detection using NLP, Machine Learning and Deep Learning

Researcher used both Machine Learning Methods and Deep learning method to get the result. Machine learning methods such as Support Vector Machine (SVM), Naive Bayes (NB) and Maximum Entropy (ME) were used which belong to the shallow structure machine learning method which are easy to implement but computationally

intensive.Researcher got to know that the advantage of machine learning-based sentiment analysis is that they have the ability to model many features. But, compared with the sentiment analysis based on emotion dictionary and machine learning, the sentiment analysis method based on depth learning has its own unique advantages and gets rid of the shackles of feature engineering. This is mainly due to the strong expression of the deep network structure, and these deep network models can utilize the semantic synthesis principle to synthesize the high-level text sentiment semantic feature vectors of low-level vectors to obtain the high-level sentiment semantic expression of the text, which effectively enhances the promotion of the model ability. Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) and Long Short-Term Memory Models were used by the researcher to obtain the following result. We can see that the maximum accuracy was achieved by the CNNs + Word2vec model.[2]

2.1 Summary of Related Work

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

Table 1 Summary of literature survey

Literature	Observation
Cecilia Ovesdotter Alm, Dan Roth and Richard Sproat[6]	The algorithm used here is sep-tune-val. The accuracy achieved was 63% .
Prof. Hardik S. Jayswal, Dhruvi D. Gosai and Himangini J. Gohil [1]	The input text was classified into different emotions by finding the emotional content from the given English text.
Douiji Yasminaa, Mousannif Hajarb and Al Moatassime	Statistical measures was used in order to compute the semantic relatedness between words of a given sentence and the target emotions.average precision of 92.75%, and 68.82% as average accuracy.

Hassana[3]	
Bo Pang, Lillian Lee and Shivakumar Vaithyanathan[5]	Naive bayes algorithm, entropy classification, and support vector machines was implemented over the unigrams. The accuracy achieved was 82% using SVM algorithm.
Kashif Khan, Sher Hayat and Muhammad Ejaz khan et al. 2016 [4]	Accuracy is 7 times more in Hybrid approach than that of other lexical approach and keyword approach. Machine learning approach is perfect algorithm for Emotion De

3. Proposed Work

The proposed system will load the trained SVM model. It will take the input text. The model will analyse the text and calculate the emotion of the input text. The System will then give a visual representation of the emotion.

3.1 System Architecture

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

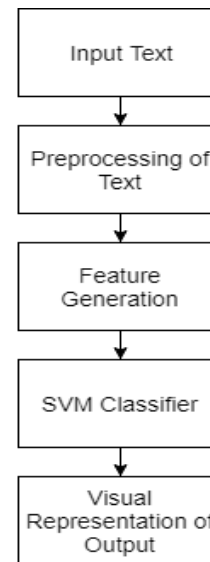
A. Input Text Block: Input from the user is taken in Hindi language.

B. Preprocessing of Text: The input text is taken and preprocessed and tokens are generated. Stop words are removed from the tokens. From the remaining tokens root words are identified. The process of stemming is done on the identified root words. The prefixes and suffixes which are associated with the root word is removed. Lemmatization is done properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word.

Algorithm:

- Preprocess the input text. Tokenize the input text. In tokenization the token of words from the sentences are formed.
- Remove stop words from the tokens.

- Identify Root Words.
- Do stemming on the identified root words. Remove the prefixes and suffixes which are associated with the root word.
- Do lemmatization with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word



Basic Architecture Of The Model

Fig. 1 Proposed system architecture

C. Feature generation: The preprocessed text is taken as the input and the features are generated in the form of Ngrams (N-gram is a contiguous sequence of n items from a given sample of text or speech. The items can be phonemes, syllables, letters, words or base pairs according to the application). Emotional Score of the ngram is calculated using the Emotion Word Corpus. Ngrams are then ordered according to the sentence structure and is added to the list of features. Summation of the scores of all the words associated with the list is taken and then returned.

Algorithm:

- Take preprocessed text as the input and generate features.
- Form ngrams from the preprocessed text. (N-gram is a contiguous sequence of n items from a given sample of text or speech. The items

can be phonemes, syllables, letters, words or base pairs according to the application).

3. Calculate the Emotional Score of the ngram using the Emotion Word Corpus.
4. Order the ngrams according to the sentence structure.
5. Add it to the list of features.
6. Sum up the scores of all the words associated with the list.
7. Return the list.

D. SVM Classifier: Decide the percentage of the dataset to be trained and then train the model using Support Vector Machine and test it against the decided percentage. Calculate the accuracy of our model and save trained model.

Algorithm:

1. Decide the percentage of the dataset to be trained.
2. Train the model using Support Vector Machine and then test it against the decided percentage.
3. Calculate the accuracy of our model
4. Save trained model.

E. Visual Representation of Output: The calculated accuracy of the trained model will be visually displayed on the screen.

3.2 Dataset and Evaluation Metrics

A dataset will be created using the sentences from the news site from the internet and social media sites. This will be done using web scraping of the particular sites and extracting sentences from them. A corpus will be created of the emotional words and emotions will be categorised. The quality of a EA system can be evaluated by comparing the emotions of the text manually to the results of the different approaches.

We use the Simple accuracy measure to check the overall accuracy of the system. This approach uses the emotions detected to the total number of emotions present in the passage.

$$accuracy = \frac{\text{Number of Correct Queries Detected}}{\text{Total No Of Queries}}$$

ACKNOWLEDGMENT

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

REFERENCES

1. Prof. Hardik S. Jayswal, Dhruvi D. Gosai and Himangini J. Gohil, "A review on emotion detection and recognition from text using natural language processing", (2018).
2. Fan Xia and Zhi Zhang, "Study of Text Emotion Analysis Based on Deep Learning", (2018).
3. Douji yasminaa, Mousannif Hajarb and Al Moatassime Hassana, "Using YouTube comments for text-based emotion recognition", (2016), a Faculty of Science and Technology, Abdelkarim Elkhatabi Street, Guéliz, Marrakesh P.C 40549, Morocco b Faculty of Semailia, Prince My Abdellah Street, Marrakesh P.C 42390, Morocco.
4. Kashif Khan, Sher Hayat and Muhammad Ejaz Khan, "Emotion Detection through Text: Survey", (2016).
5. Bo Pang, Lillian Lee and Shivakumar Vaithyanathan : "Thumbs up? Sentiment Classification using Machine Learning Techniques", (2002).
6. Cecilia Ovesdotter Alm, Dan Roth and Richard Sproat, "Emotions from text: machine learning for text-based emotion prediction", (2005).

Intelligent Agriculture Greenhouse Environment Monitoring System Based on Internet of Things Technology

CONFERENCE ON TECHNOLOGIES FOR FUTURE CITIES (CTFC) 2019

PROF. PAYEL THAKUR

ARADHANA POTTETH, SHWETA PURUSHOTHAMAN, JIDNYESHA TAKLE, ROHINI BRIDGITTE STANLY

Abstract:

Nowadays, technology is being used in our daily life. If agriculture is combined with automation, it will reduce manual hard work to a great extent. IOT (Internet of Things) technology was developed for connecting a billion of devices to an Internet. This technology has become very useful in agricultural modernization. A huge amount of information is transferred between the electronic devices. It is a new way to interact between device and people. We will use CC2530 chip as the core which is based on ZigBee technology.[3] This chip will be connected to Raspberry pi. Sensor nodes will be connected to CC2530 chip. The system will be made to control temperature, humidity, moisture and light inside the greenhouse. The sensor nodes will sense the parameters inside the greenhouse and will provide notification to the user if necessary. User will control the parameters using Android application accordingly.

Keywords:

Internet of Things, Raspberry Pi, , CC2530, CC2530F256, Zigbee technology

Submitted on: 23 October 2018

Revised on:

Accepted on:

*Corresponding Author Email: rohinistanly@gmail.com

Phone: 8828753911

I. INTRODUCTION

This paper introduces a kind of greenhouse monitoring system which is constructed based on Zigbee technology.

The main objective of this project is to build greenhouse with automatic monitoring and controlling system, i.e constantly monitor and control environmental conditions in greenhouse. It focuses on saving water, increasing efficiency and reducing the environmental impacts on production of plants. The user can see the atmospheric conditions of the greenhouse plants on android app and control the greenhouse from faraway places. It is to increase the production of food stuffs and to save water, power etc.

Principle rule of the system is to control the present environmental conditions of the Greenhouse using sensors and chips. For IoT based system, the sensors and the chips will be controlled by Raspberry Pi 3. The chip for controlling sensors will be CC2530 more specifically CC2530F256 which provides a robust and complete ZigBee solution. The entire system will be managed manually using Android application.

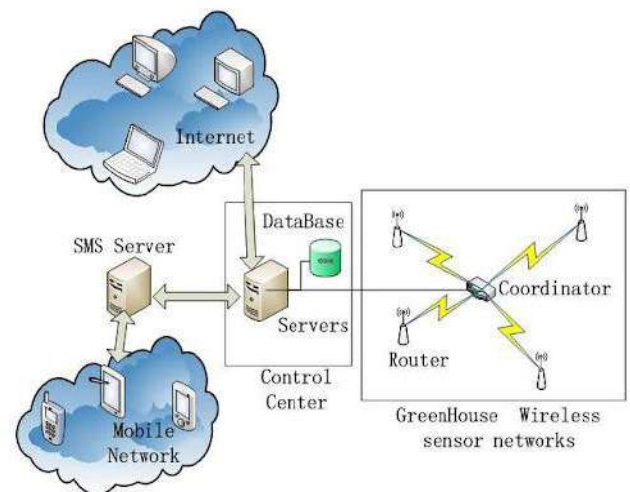


Fig. 1 Principle of the system[8]

II. RELATED WORK

We referred various research papers. Out of the ten papers, six of them were based on Internet of Things technology. While two were based on Android platform. The remaining two depends on Micaz motes and embedded Web server technology respectively.

“Liu Dan Cao Xin Haung Chongwei Ji Liangliang”[3] et al. proposed greenhouse monitoring system by considering CC2530 chip as

its core in WSN, the system is made up of front end data acquisition, data processing, data transmission and data reception. The ambient temperature is real time processed by temperature sensor, processed data is send to intermediate node aggregates all data and sends it to PC through serial port; at the same time, staff may view and send it.

To meet the needs of remote monitoring of greenhouse environment parameters, combined with embedded technology and 3G communication technology, a scheme of greenhouse environment parameter information real-time monitoring and control based on the Android phone platform is proposed in the paper proposed by “Li Zhang, Congcong Li” et. al.[6]

Integrating web and embedded technology, “Gao Junxianga” et. al. proposes a design of monitor system for greenhouse based on embedded web server and wireless sensor network. A tiered architecture monitor system is discussed firstly, and then detailed design of the system is given including hardware and software of embedded web server and wireless sensor network. The embedding way of web server in the device enable the embedded devices to be connected to the Internet and also enable users to access, control and manage the embedded devices using a standard web browser over the Internet without restrict of time and space.[7]

“Mustafa Alper Akkas” et. al. presents a WSN prototype consisting of MicaZ nodes which are used to measure greenhouses’ temperature, light, pressure and humidity. Measurement data have been shared with the help of IoT. With this system farmers can control their greenhouse from their mobile phones or computers which have internet connection.[5]

III. METHODOLOGY

The major components are Raspberry pi, GSM, a block consisting of factors such as temperature, humidity, light intensity and soil moisture and a block of actuators including fan, spray, light source and motors.

The Raspberry Pi 3 is a small single-board computer with 64 bit quad core processor, on-board WiFi, Bluetooth and USB boot capabilities. By adding a keyboard, mouse, display, power supply, micro SD card with installed Linux Distribution, it

will act as a fully fledged computer that can run applications from word processors and spreadsheets to games.

The network topology model of Zigbee is satellite. The Zigbee coordinator is the organizer of Zigbee network. It receives the wireless sensor nodes information and sends the information to the gateway through the serial port. Zigbee is a wireless technology developed as an open global standard to address the unique needs of low-cost, low-power wireless IoT networks. The Zigbee standard operates on the IEEE 802.15.4 physical radio specification and operates in unlicensed bands including 2.4 GHz, 900 MHz and 868 MHz. Zigbee devices can transmit data over long distances by passing data through a mesh network of intermediate devices to reach more distant ones. Zigbee has a defined rate of 250 kbit/s, best suited for intermittent data transmissions from a sensor or input device.

The sensor will sense the parameters such as temperature, humidity, light intensity and soil moisture present inside the greenhouse. If the parameters deviates from the threshold value, the user will get a notification in his cell phone via Android application.

An Android application is a software application running on the Android platform. Because the Android platform is built for mobile devices, a typical Android app is designed for a smartphone or a tablet PC running on the Android OS. The user will be able to control the greenhouse via installed actuators. Actuators include fan, sprinkler, light source such as LEDs and motor.

There are various applications of intelligent agriculture greenhouse environment monitoring system based on Internet of Things (IoT). The project is inclined towards a number of social applications. Various applications include Horticulture, Precision agriculture (PA) or Site Specific Crop Management (SSCM), Floriculture or flower farming, Greenhouse automation, Crop management, Smart farming, End-to-end farm management systems.

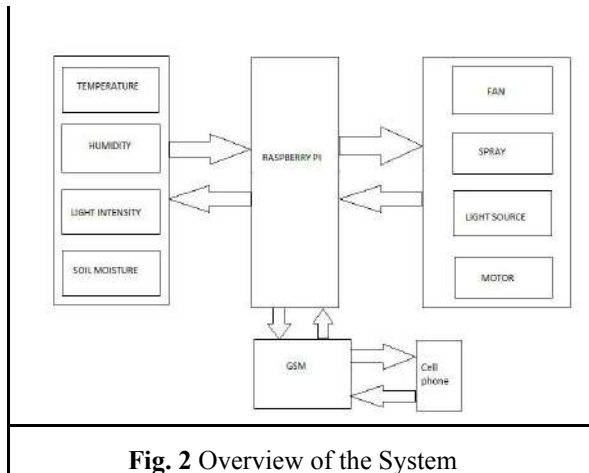


Fig. 2 Overview of the System
IV. EXPERIMENTATION

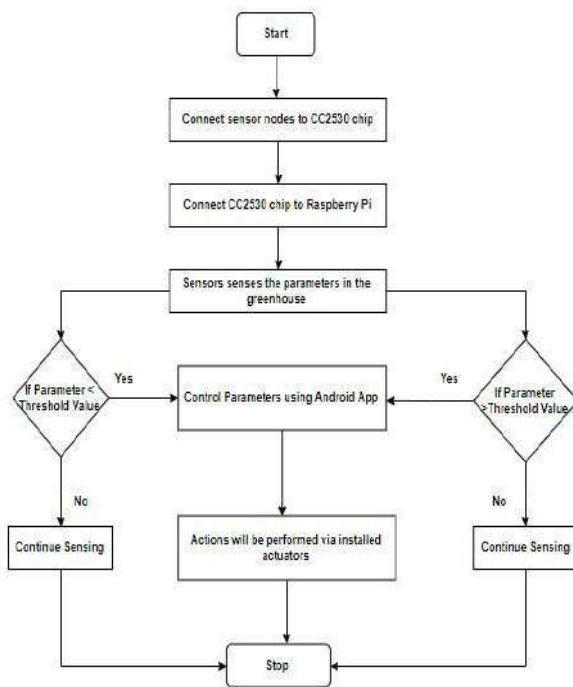


Fig. 3 Flowchart

The system is made up of front-end data acquisition, data processing, data transmission and data reception. The temperature is processed in real time by the temperature sensor of data terminal node. Processed data is sent to the intermediate node via wireless network. The intermediate node aggregates all data and sends it to the PC, at the same time, staff may view, analyse or store data by the PC that provides real-time data for agriculture greenhouse, fans and other temperature control equipment, to achieve automatic temperature control. Connect sensor nodes to CC2530 chip.

Connect CC2530 chip to Raspberry pi. Sensor senses the parameters inside the greenhouse. If parameters exceed the threshold value, control the parameters using Android App via installed actuators. Else, continue sensing. Similarly, if parameters falls behind the threshold value, control the parameters using Android app via installed actuators. Else, continue sensing.

V. RESULTS AND DISCUSSION

There are various parameters present inside the greenhouse. Attributes such as temperature, humidity, light intensity and soil moisture are received via sensors. The sensors are there inside the greenhouse. The inputs for Android application are user controlled parameters and threshold values. Let us discuss the output. For the android application, various actuators such as fan, spray, light source and motor can be considered as the output.

The table given below represents sample dataset. The sample dataset consists of parameters and their corresponding threshold values. The dataset includes temperature, humidity, moisture and light intensity. The threshold value for temperature is 77F whereas the threshold for humidity is 35%. The standard values for moisture and light intensity are 32% and 33.8% respectively.

Dataset	Threshold
Air Temp (F)	77
Humidity (%)	35
Moisture (%)	32
Light intensity (%)	33.8

Table 1. Sample Dataset

VI. CONCLUSION

We are designing an Android app which can be easily installed in any platform. As it is an app, we can use it anytime, anywhere. This way, mobility can be achieved. The low cost, low power wireless

Zigbee technology applies in greenhouse monitoring system. The system realizes the remote intelligent control to the room equipment through Internet. It improves the operational efficiency and system application flexibility by using the wireless sensor network instead of the traditional wired network, and at the same time reduces the manpower cost. The environment data of the greenhouse can transfer reliably, and the control instruction sent timely. This design realizes remote intelligent monitoring and control of greenhouse, and is helpful to farms to scientific and rational planting crops.



Things", China Agricultural University, Beijing, China.

9. S. Muthupavithran, S. Akash, P. Ranjithkumar (2016), "Greenhouse Monitoring using Internet of Things", Vellammal Engineering College, Chennai, India.
10. Varsha Modani, Ravindra Patil, Pooja Puri, Niraj Kapse (2017), "IoT Based Greenhouse Monitoring System :Technical Review", IRJET, India.

VI. REFERENCES

1. LIU Dan, Wan hongli , Zhang Mengya , Xiang Jianqiu (2017), "Intelligent Agriculture Greenhouse Environment Monitoring System Based on the Android Platform", IEEE, Dalian, China.
2. Ravi Kishore Kodali, Vishal Jain and Sumit Karagwal (2016), "IoT based Smart Greenhouse", IEEE, Warangal, India.
3. LIU Dan, Cao Xin, Huang Chongwei, Ji Liangliang (2015), "Intelligent Agriculture Greenhouse Monitoring System Based on IOT Technology", IEEE, Dalian, China.
4. Zaidon Faisal Shenan, Ali Fadhil Marhoon, Abbas A. Jasim. (2017), "IoT Based Greenhouse Monitoring and Control System", Basrah Journal of Science, Basrah, Iraq.
5. Mustafa Alper Akkas, Radosveta Sokullub (2017), "An IoT-based greenhouse monitoring system with Micaz motes", PCS, izmir, Turkey.
6. Li Zhang, Congcong Li, Yushen Jia, Zhigang Xiao (2015), "Design of Greenhouse Environment Remote Monitoring System Based on Android Platform", AIDIC, China.
7. Gao Junxianga, Du Haiqingb (2011), "Design of Greenhouse Surveillance System Based on Embedded Web Server Technology", PCS, Wuhan, China.
8. Shaoling Li, Yu Han, Ge Li, Man Zhang, Lei Zhang, Qin Ma (2011), "Design and Implementation of Agricultural Greenhouse Environment Remote Monitoring System Based on Internet of

i. Author Biographical Statements

	<p>Prof. Payel Thakur Assistant Professor Pillai College of Engineering</p>
	<p>Aradhana Potteth BE Computer Pillai College of Engineering</p>

	<p>Shweta Purushothaman BE Computer Pillai College of Engineering</p>
	<p>Jidnyesha Manohar Takle BE Computer Pillai College of Engineering</p>
	<p>Rohini Bridgitte Stanly BE Computer Pillai College of Engineering</p>

Human Posture Recognition and movement analysis Using Convolution Neural Network

Prathamesh N. Patade, Tejas Desai, Harman Singh Bath and Prof. Prakash Bhise,

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

Abstract—Human posture refers to the arrangement of the body and its limbs. Human posture recognition is a computer vision problem. Traditional algorithms have to deal with very large number of feasible human postures, large changes in human appearance, part occlusions and also the presence of multiple people within close proximity to each other results in figure overlapping issues. We present an approach to effectively detect the posture of multiple people in a video feed. The approach uses a Convolution Neural Network (CNN) to obtain confidence maps for body parts and a non-parametric representation called as Part Affinity Fields (PAFs) for part association. A greedy inference parses the confidence maps and PAFs to output 2D key points for all people in the image. Human posture recognition is gaining increasing attention in the field of computer vision due to its promising applications in the areas of personal health care, environmental awareness, human-computer-interaction, sports monitoring and surveillance systems. Progress in this area will lead to wide applications such as human tracking, action recognition, video analysis, anomaly detection and a lot more.

Keywords—Part Affinity Field (PAFs), Convolution Neural Network (CNN), Deep Learning

1. Introduction

Body posture refers to the position or orientation of a person's body. Different positions are defined by different names, all are which are different body postures. These identified postures are use in ergonomics are for defining the demands of activities performed by humans.

Body postures are defined by identifying the part of the body being positioned and its positioning. For example, straight, twisted and stooped are all defined as different back postures. As for when defining arm postures, terms such as below shoulder or overhead are used. Leg postural terms include sitting, standing, crawling or walking.

Image processing is a method to convert an image into digital form and perform some operations on it,

in order to get an enhanced image or to extract some useful information from it. It is a type of signal dispensation in which input is image, like video frame or photograph and output may be image or characteristics associated with that image. Usually Image Processing system includes treating images as two-dimensional signals while applying already set signal processing methods to them.

It is among rapidly growing technologies today, with its applications in various aspects of a business. Image Processing forms core research area within engineering and computer science disciplines too.

2. Literature Survey

A. Pose Estimation using PM-ensemble model

This approach presents a PM-ensemble (PME) model to infer body configurations by modelling the interdependence among the responses of PM models. The model training process consists of three stages. At stage 1, the training samples are partitioned into subsets based on their similarity in a pose space. At stage 2, each PM model is trained using training samples from each cluster. At stage 3, learning of the PME model to incorporate all the responses to make the final estimation. Given an input image, the learned PM models is used to localize body joint positions independently.

B. Articulated Pose Estimation via Inference Model

This presents a method for articulated human pose estimation that builds off the hierarchical inference machine originally used for scene parsing. Conceptually, the presented method, which is referred to as a Pose Machine, is a sequential prediction algorithm that emulates the mechanics of message passing to predict a confidence for each variable (part), iteratively improving its estimates in each stage. The inference machine architecture is particularly suited to tackle the main challenges in pose estimation. First, it incorporates richer

interactions among multiple variables at a time, reducing errors such as double counting. Second, it learns an expressive spatial model directly from the data without the need for specifying the parametric form of the potential functions. Third, its modular architecture allows the use of high capacity predictors which are better suited to deal with the highly multi-modal appearance of each part. Inspired by recent work that has demonstrated the importance of conditioning finer part detection on the detection of larger composite parts in order to improve localization, this incorporates these multi-scale cues in the framework by also modelling a hierarchy of parts.

C. Joint Subset Partition and Labeling

As a principled solution for multi person pose estimation model is proposed that jointly estimates poses of all people present in an image by minimizing a joint objective. The formulation is based on partitioning and labelling an initial pool of body part candidates into subsets that correspond to sets of mutually consistent body-part candidates and abide to mutual consistency and exclusion constraints. The formulation is able to deal with an unknown number of people, and also infers this number by linking part hypotheses. The formulation allows to either deactivate or merge part hypotheses in the initial set of part candidates hence effectively performing non-maximum suppression (NMS). In contrast to NMS performed on individual part candidates, the model incorporates evidence from all other parts making the process more reliable. The problem is cast in the form of an Integer Linear Program (ILP). Although the problem is NP-hard, the ILP formulation facilitates the computation of bounds and feasible solutions with a certified optimality gap. This paper makes the following contributions. The main contribution is the derivation of a joint detection and pose estimation formulation cast as an integer linear program. Further, two CNN variants are proposed to generate representative sets of body part candidates. These, combined with the model, obtain state-of-the-art results for both single-person and multi-person pose estimation on different datasets.

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.

2.1 Summary of Related Work

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

Table 1 Summary of literature survey

SN	Paper	Advantages and Disadvantages
1.	V. Ramakrishna, Pose machines: Articulated pose estimation via inference machines. (2014)	<u>Disadvantages:</u> Top-down approach is proportional to the number of people: for each detection, a single-person pose estimator is run, and the more people there are, the greater the computational cost.
2.	1) L. Pishchulin, E. Insafutdinov. Joint subset partition and labeling for multi person pose estimation. (2016) 2) E. Insafutdinov, B. Andres. Deepcut: A deeper, stronger, and faster multiperson pose estimation model.(2016)	<u>Disadvantage:</u> Costly Global inference and solving integer linear programming problem over a fully connected graph is a NP-hard problem and the average processing time is on the order of hours.
3.	K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition.(2016)	<u>Advantages:</u> improved runtime. <u>Disadvantages:</u> Method still takes several minutes per image, with a limit on the number of part proposals. Pairwise representations used in are difficult to regress precisely and thus a separate logistic regression is required.

4.	Zhe Cao, Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields (2017)	<u>Disadvantage:</u> Frame-by-frame extraction of images is computationally expensive to provide real-time results.
----	---	---

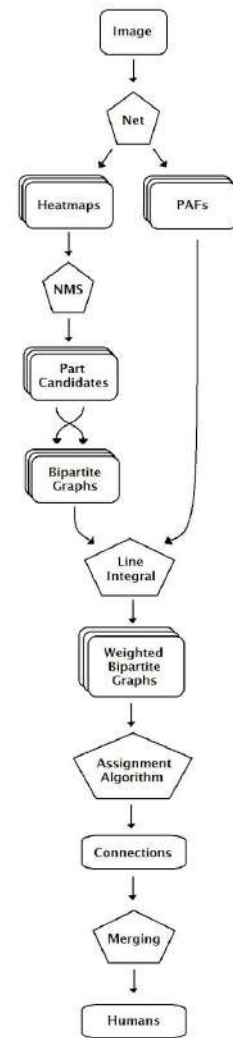


Fig. 1 Proposed system architecture

3. Proposed Work

3.1 System Architecture

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

1. Generate heatmaps and PAFs (Part Affinity fields)

The image is passed through the two-branch multi-stage CNN. Each stage in the first branch predicts confidence maps S^t , and each stage in the second branch predicts PAFs L^t . After each stage, the predictions from the two branches, along with the image features, are concatenated for the next stage.

$$S^t = \rho^t(\mathbf{F}, S^{t-1}, L^{t-1}), \forall t \geq 2, \quad (1)$$

$$L^t = \phi^t(\mathbf{F}, S^{t-1}, L^{t-1}), \forall t \geq 2, \quad (2)$$

Where ρ^t and ϕ^t are the CNNs for inference at Stage t .

Two L2 loss functions are applied at the end of each stage, one at each branch find loss between estimated predictions and the ground truth maps and fields. L2 loss is selected between the estimated predictions and ground truths for both sub-networks.

$$f_S^t = \sum_{j=1}^J \sum_{\mathbf{p}} W(\mathbf{p}) \cdot \|S_j^t(\mathbf{p}) - S_j^*(\mathbf{p})\|_2^2,$$

$$f_L^t = \sum_{c=1}^C \sum_{\mathbf{p}} W(\mathbf{p}) \cdot \|L_c^t(\mathbf{p}) - L_c^*(\mathbf{p})\|_2^2,$$

Where S_j^* is the ground truth part confidence map, L_c^* is the ground truth part affinity vector field, W is a binary mask with $W(\mathbf{p})=0$ when the annotation is missing at an image location \mathbf{p} .

The overall objective is

$$f = \sum_{t=1}^T (f_S^t + f_L^t).$$

In the loss of both branches, W is provided to distinguish between locations missing the annotation or not. The final loss is simply adding two together.

2. Generation of confidence map

The heatmaps are passed through NMS algorithm. The ground truth confidence map is generated using the raw figure and 2D annotations for each part. For every location on the map, the confidence value should be

$$S_j^*(\mathbf{p}) = \max_k S_{j,k}^*(\mathbf{p}).$$

We apply a non-maximum suppression (NMS) algorithm to get those peaks.

1. Start in the first pixel of the heatmap.
2. Surround the pixel with a window of side 5 and find the maximum value in that area.
3. Substitute the value of the center pixel for that maximum
4. Slide the window one pixel and repeat these steps after we've covered the entire heatmap.

1. Compare the result with the original heatmap. Those pixels staying with the same value are the peaks we are looking for. *Suppress* the other pixels setting them with a value of 0.

After all the process, the non-zero pixels denote the location of the part candidates.

3. Part Association

3.1 Bipartite Graph

From step 2 we get the candidate for each one of the body parts. Using these part candidates we have to create the complete bipartite graph, where vertices are the part candidates and the edges are the connection candidates. Let's say we have a set of neck candidates and a set of right hip candidates and the edges are the connection and a set of right hip candidates. Then the bipartite graph is created by connecting each node in one set to every other node in the second set. Right connection between vertices is found out by the assignment problem.

3.2 Line Integral

To solve assignment problem each edge on the graph should have a weight. We will compute the line integral along the segment connecting each couple of part candidates, over the corresponding PAFs (x and y) for that pair. Line integral measures the effect of a give field along a given curve.

$$E = \int_{u=0}^{u=1} L_c(\mathbf{p}(u)) \cdot \frac{\mathbf{d}_{j_2} - \mathbf{d}_{j_1}}{\|\mathbf{d}_{j_2} - \mathbf{d}_{j_1}\|_2} du,$$

3.3 Assignment

The weighted bipartite graph shows all possible connection between candidate of two parts and holds score for every connection.

1. Sort each possible connection by its score.
2. The connection with the highest score is indeed a final connection.
3. Move to next possible connection. If no parts of this connection have been assigned to a final connection before, this is a final connection.
4. Repeat the step 3 until we are done.

3.4 Merging

In the final step detected connections are transform into final skeletons. At first, every connection belongs to a different human. This way, we have the

same number of humans as connections we have detected.

Let *Humans* be a collection of sets $\{H_1, H_2, \dots, H_k\}$. Each one of these sets — that is, each human — contains, at first, two parts (a pair). And let's describe a part as a tuple of an index, a coordinate in the 'x' direction and a coordinate in the 'y' direction.

$$\text{Humans} = \{H_1, H_2, \dots, H_k\}$$

where

$$k := \text{number of final connections}$$

$$H_i = \{(m_{idx}, m_x, m_y), (n_{idx}, n_x, n_y)\}$$

Here comes the merging: if humans H_1 and H_2 share a part index with the same coordinates, they are sharing the same part! H_1 and H_2 are, therefore, the same humans. So we merge both sets into H_1 and remove H_2 .

```

if  $H_1 \cap H_2 \neq \emptyset$ 
then
   $H_1 = H_1 \cup H_2$ 
  delete( $H_2$ )

```

3.2.3 Posture comparison with predefined pose templates.

If the input pose matches the model pose essentially comes down to checking if the two poses have the same shape. This can be treated as a Procrustes Problem:

To compare the shapes of two objects, the objects must be first optimally “superimposed”. Procrustes superimposition (PS) is performed by optimally translating, rotating and uniformly scaling the objects. In other words, both the placement in space and the size of the objects are freely adjusted. The aim is to obtain a similar placement and size, by minimizing a measure of shape difference called the Procrustes distance between the objects.

The difference between the shape of two objects can be evaluated only after “superimposing” the two objects by translating, scaling and optimally rotating them as explained above. Consider the perfect case (identical shapes), after PS, the objects will perfectly coincide. Of course, e.g. due to different body proportions, there is no way a person will succeed in perfectly copying the model pose. There is a need for some thresholding; If all corresponding points are approximately adjacent, we can conclude the two

poses match. From the moment a distance exceeds a certain threshold, they're different.

We're looking for the combination of a translation, scaling and rotation that best transforms the input pose onto the model pose. From Linear Algebra, we know this combination of operations is wrapped in a linear transformation, more precisely an affine transformation (composition of linear map and a translation).

The properties:

- Lines maps to lines
- Parallel lines remain parallel
- Origin does not necessarily map to origin
- Ratios are preserved

3 Requirement Analysis

The implementation detail is given in this section.

3.1 Software

Table 3.1 Software details

Operating System	Linux
Languages used	Python, PHP, HTML, Flask
Framework	Keras, Tensorflow

3.2 Hardware

Table 3.2 Hardware details

Processor	3.5 GHz Intel i7
HDD	180 GB
RAM	16 GB or more

GPU	4 GB or more VRAM NVIDIA GPU preferred.
-----	--

3.3 Dataset and Parameters

1. MPII Human Pose dataset

MPII Human Pose dataset is a state-of-the-art benchmark for evaluation of articulated human pose estimation. The dataset includes around **25K images** containing over **40K people** with annotated body joints. The images were systematically collected using an established taxonomy of every day human activities. Overall the dataset covers **410 human activities** and each image is provided with an activity label. Each image was extracted from a YouTube video and provided with preceding and following un-annotated frames. In addition, for the test set we obtained richer annotations including body part occlusions and 3D torso and head orientations.

ACKNOWLEDGMENT

It is our privilege to express our sincerest regards to our supervisor Prof. Prakash Bhise 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. Madhumita Chatterjee and our Principal Dr. Sandeep M. Joshi for encouraging and allowing us to presenting this work.

REFERENCES

[1] Zhe Cao, Tomas Simon and Shih-En Wei and Yaser Sheikh. “*Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields.*” In CVPR, 2017.

[2] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. “*Deepcut: A deeper, stronger, and faster multiperson pose estimation model.*” In ECCV, 2016.

[3] A. Newell, K. Yang, and J. Deng. “*Stacked hourglass networks for human pose estimation.*” In ECCV, 2016.

[4] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. “*Convolutional pose machines.*” In CVPR, 2016.

[5] L. Pishchulin, A. Jain, M. Andriluka, T. Thormahlen, and B. Schiele. “*Articulated people detection and pose estimation: Reshaping the future.*” In CVPR, 2012.

[6] G. Gkioxari, B. Hariharan, R. Girshick, and J. Malik. “*Using k-poselets for detecting people and localizing their key points.*” In CVPR, 2014.

[7] MSCOCO keypoint dataset. <http://mscoco.org/dataset/#keypoints-eval>

[8] MPII Human Pose Dataset. <http://human-pose.mpi-inf.mpg.de>

[9] Yuki Kawanab, Norimichi Ukitaa, Jia-Bin Huangc , Ming-Hsuan Yang. “*Ensemble Convolutional Neural Networks for Pose Estimation.*” CVIU 2014

[10] Varun Ramakrishna, Daniel Munoz, Martial Hebert, J. Andrew Bagnell, and Yaser Sheikh. “*Pose Machines: Articulated Pose Estimation via Inference Machines*” CVIU 2015

[11] K. Simonyan and A. Zisserman. “*Very deep convolutional networks for large-scale image recognition.*” In ICLR, 2015.

Proposed Voice Based Notice Board Using Android

Amritha Sajeev, Raksha Bondanthila, Deeksha Kumbla, Rohit Nair, Prof. Gayatri Hegde

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

Abstract— Notice board is a primary thing in any institution/organization or public utility places like bus stations, railway stations and parks. But sticking various notices day-to-day is a difficult process. The Notice board is a common display for effective mode of providing information to the people, but this is not easy for updating the messages instantly. This project deals about an advanced Hi-Tech wireless Notice Board. This system is enhanced to display the latest information through an Android application of smart phones or tablet. The proposed notice board is a multi user password-protected SMS based system fabricated with an LCD. The communication and information transfer between the authentic user and the LCD display unit is done via GSM to ensure remote display facilities, so any notice can be displayed on the electronic board from the user's mobile SMS from distant places. To ensure system flexibility, a multi user noticing and displaying system has been implemented in the system which can display several notices simultaneously. In addition, the user also can print any notice which is of concern to them. The total system is designed with simple logic with a robust algorithm and fabricated with a PIC midrange microcontroller, LCD, GSM module and other commercially available electronic devices to ensure efficiency and reliability with less cost. Voice based notice board using android can be used in transportation areas, stock markets and trades, educational institutes and food courts.

Keywords— notice board, electronic notice, android application, GSM, LCD; microcontroller.

1. Introduction

A notice Board is a place where an authenticated authority can leave public messages to advertise things, announce events or provide information of general concern for any important issue. But some shortcomings make this analog notice board unpopular in general. Ascribed personnel are always needed to change any notice or originate a new one. Also, multiple people gather, struggle and cluster in front of a single traditional notice board for information in case of any urgent notification. Sometimes malicious intentions of any persons can manipulate, remove or perish paper notices attached in a board, leaving other people uninformed. If the boards are placed in busy places, e.g. near entrances or exit points, then a busy person does not get enough time and scope to access and read all the informations posted on a notice board. It become more problematic when no digital printout is

possible. One other disadvantage is that these traditional boards often get dirty, having wear and tear on notices and an unorganized pattern, which make the notice board quite inconvenient for users. There is also an unregulated display of information, difficulty in storage.

Main concept behind Voice operated Electronic notice board using display is to show messages and to control them by using our own voice. We have already seen GSM primarily based Electronic board, but speech controlled board has extra advantage of simple use. While the user sends the message from the Android application device, it is received and retrieved by the Bluetooth device at the display unit. Voice recognition is finished within the automaton application. User needs to install this automaton application. Bluetooth wireless technology is becoming a popular standard in the communication arena, and it is one of the fastest growing fields in the wireless technologies. Bluetooth technology handles the wireless part of the communication channel; it is used in this project to transmit and receive data wirelessly between devices. While a phone is simply more than a phone these days, it is a smartphone the number of applications being built on a wide range of platforms for smart phone is astounding. Speech synthesis is the artificial production of human speech. A computer system used for this purpose is called a speech synthesizer, and can be implemented in software or hardware products. A text-to-speech (TTS) system converts normal language text into speech. Now a days GSM modem based notice boards are also in use but they require router in which cable connections are done which make it complex

2. Literature Survey

A. Android Phone Speech Recognition Sensed Voice Operated Notice Board Display: This project by Sanjeev Singh is an implementation of the idea of wireless communication between a mobile phone and an Arduino. The display unit consists of LED display that is interfaced with arduino. Bluetooth is an open wireless protocol for exchanging data over short distances from fixed and mobile devices, creating Personal Area Networks (PANs). It was originally conceived as a wireless alternative to

RS232 data cables. It can connect several devices, overcoming problems of synchronization[1].

B. Raspberry Pi Based Speech Recognition Sensed Smart Notice Board Display: The proposed system consists of android phone section and a receiver section. Android phone section consists of an android mobile phone in which the announcer speaks through our own developed open source speech to text application. To transfer the information, he/ she need to speak out the message through the android phone, which is provided with internet facility. A speech to text mobile application is used to convert the spoken voice message into the text message. The converted message is then transferred to the receiver section via Email. Thus the text message will be sent to the desired email which is predefined. A HD display is connected to the Raspberry pi via the High Definition Multimedia Interface port in it. The Raspberry pi continuously checks the particular email and on receiving one, it opens the message and save it as a HTML text file in the memory card provided in the Raspberry pi. Application Programmable Interface is provided to mark the emails which are already been read. Thus it identifies the new notifications and is displayed on the HD display. The display is programmed in such a way that it appears like a blank screen and displays the notifications when they arrive. The text message will be displayed within 30 seconds. [2].

C. Voice Based Notice Board: Amit Zore uses either Bluetooth or Wi-Fi based wireless serial data communication in displaying messages on a remote digital notice board. Android based Application programs available for Bluetooth and Wi-Fi communication for personal digital assistant (PDA) devices are used for transmitting the alpha-numeric text messages. Using the Bluetooth or Wi-Fi based serial data communication technique, the corresponding transceiver module has been interfaced with Wireless notice board at the receiver end. For this purpose, a low cost wireless notice board is programmed to receive alphanumeric text messages in any of the above selected communication modes. The proposed system will help in reducing the human effort.[3].

D. Display Message On Notice Board Using GSM: It presents an SMS based notice board incorporating the widely used GSM to facilitate the communication of

displaying message on notice board via user's mobile phone. Its operation is based on microcontroller ATMEGA32 programmed in assembly language. A SIM300 GSM modem with a SIM card is interfaced to the ports of the microcontroller with the help of AT commands. When the user sends a SMS via a registered number from his mobile phone, it is received by SIM300 GSM modem at the receiver's end. SIM300 is duly interfaced through a level shifter IC MAX32 to the microcontroller. The messaged is thus fetched into the microcontroller. It is further displayed on an electronic notice board which is equipped with LCD display interfaced to microprocessor powered by a regulated power supply from mains supply of 230 volts ac[4].

2.1 Summary of Related Work

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

Table 1 Summary of literature survey

Literature	Advantages	Disadvantages
Sanjeev Singh et al. 2017 []	Low Cost Less Complex	Doesn't overcome distant connectivity
Neenu Ann George et al. 2016 []	Flexible Reliable	Inefficient mail checking
Amit Zore et al. 2017 [3]	Low Cost Less Complex	RF lacks distant connectivity
Forum Kamdar et al. 2018 [4]	Multiple users and Distant Connectivity possible	Multiple login results in complexity regarding to proper access

3. Proposed Work

Controlling the computer mouse using the eyes movement requires a fast and effective algorithm, that's brought us to decrease the running time of the tool to the minimum by dividing the operation into few steps and using a In view of the above it will be apparent that, there exists a need of electronic notice board that enables efficient way to the user for displaying notice. By considering

increasing compactness of electronic systems, there is a need of embedding two or more systems together. This project is an implementation of the idea of wireless communication between a mobile phone and an AVR controller. In this project work, we are supposed to design an embedded system which consists of display unit, printer and audio device using wireless technology. The display unit consists of any type of display that can be interfaced with microcontroller. Wireless printer is used for printing application. Audio device is speaker which is controlled by microcontroller through Text-To-Speech (TTS) convertor. GSM technology is specially used for SMS applications.

3.1 System Architecture

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

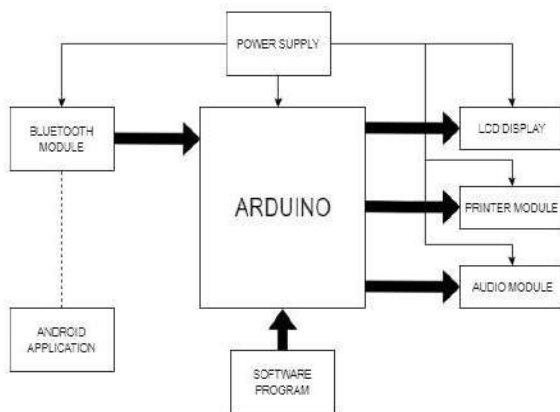


Fig. 1 Proposed system architecture

A. Bluetooth Module: A Bluetooth module is usually a hardware component that provides a wireless product to work with the computer; or in some cases, the bluetooth may be an accessory or peripheral, or a wireless headphone. or other product (such as cellphones can use) If the computer (is this computer related?) has hardware support to.

B. Android Application: an Android application that is capable of performing the following

Functions:

Convert voice data to text

Send this text over to microcontroller via Bluetooth for displaying on notice board

Play the message from the audio device

Send the message as SMS to anybody

C. Arduino: Arduino is an open-source computer hardware and software company, project and user community that designs and manufactures kits for building digital devices and interactive objects that can sense and control the physical world. An Arduino board consists of complementary components that facilitate programming and incorporation into other circuits. An important aspect of the Arduino is its standard connectors, which lets users connect the CPU board to a variety of interchangeable add-on modules known as shields. An Arduino's microcontroller is also pre-programmed with a boot loader that simplifies uploading of programs to the on-chip flash memory, compared with other devices that typically need an external programmer. At a conceptual level, when using the Arduino software stack, all boards are programmed over an RS-232 serial connection. Serial Arduino boards contain a level shifter circuit to convert between RS232-level and TTL-level signals.

D. LCD Display: This is the first interfacing example for the Parallel Port. We will start with something simple. This example doesn't use the Bi-directional feature found on newer ports, thus it should work with most, if not all Parallel Ports. It however doesn't show the use of the Status Port as an input for a 16 Character x 2 Line LCD Module to the Parallel Port. These LCD Modules are very common these days, and are quite simple to work with, as all the logic required running them is on board.

E. Printer Module: Wireless printers refers to printers in which a radio frequency (RF) or infrared light (IR) interface connects the printer to the network, a controlling PC, a handheld computer, or both. Wireless printers come in different sizes and shapes, from full-featured stationary models to small. The wireless interface eliminates the need for cables, eradicating a potential failure point and the subsequent repair or replacement cost, while providing a safer and more space-efficient work area. The AVR controller process the data and send it to the display unit, printer and audio device.

F. Audio Module: Audio device is speaker which is controlled by microcontroller through Text-To-Speech (TTS) convertor. The programming of AVR controller will be done in assembly language. Text-To-Speech (TTS) convertor is connected serially to the AVR and after that convertor amplifier is connected to amplify the audio signal. If the message is very much important then audio device will announce it.

3 Requirement Analysis

The implementation detail is given in this section.

3.1 Software

Website:

Website is used to download the application into the Android phones by scanning the QR code.

Languages: HTML, CSS, PHP, JavaScript,

Database: MySQL

Software: Visual Studio, XAMPP

Android Application:

Android Application used in mobiles/tablets is used to send the voice message to display on the notice board

Language: Java

Software: Android Studio

3.2 Hardware

Bluetooth Module: Bluetooth Module acts as a bridge between the application and Arduino module.

Arduino: It supports LCD display, audio and printer.

LCD Display: A 20*4 LCD Display is used to display the message which was sent in the form of voice whereas a printer and an audio device will be connected to the Arduino for hardcopy and voice based output.

4. Applications

- It can be used in colleges, schools, bus stands and railway stations.
- Industries where we can control machines by just saying instructions.
- It can be used in malls & highways for advertisement purpose.
- It can be used in educational premises like schools, colleges, university campuses. It can be used to display information like exam schedule, notice, event notification and exam result announcement.

ACKNOWLEDGMENT

It is our privilege to express our sincerest regards to our supervisor Dr. Satishkumar Varma for the valuable inputs, able guidance, encouragement, whole-hearted cooperation and constructive criticism throughout the duration of this work. We would also take this opportunity to thank our mentor Prof Gayatri Hegde for her guidance in selecting this project and also for providing us with all the details needed for the project. We are also grateful to our HOD Dr. Madhumita Chatterjee for extending her help directly and indirectly through various channels in our project 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. *Sanjeev Singh, Sharad Yadav, Rajat Agarwal, Shubham Bansal*, "Android Phone Speech Recognition Sensed Voice Operated Notice Board Display," IJARCCCE, Vol-6, no. 4, pp. 2278-1021, Abbrev. April, 2017.
2. *Neenu Ann George, Prabitha. P, Priyanka.A.K, Ershad.S.B*, "Raspberry Pi Based Speech Recognition Sensed Smart Notice Board Display," IJSRD, no. 12, vol. 3, pp. 2321-0613, Nov. 2016.
3. *Prof. Amit Zore, Ms. Snehal Langhe, Ms. Sadaf Jahagiradar, Ms. Jyoti Bhosale, Ms. Pooja Pawar*, "Voice Based Notice Board," IERJ, no. 4, vol. 2, pp. 4153-4155, Nov. 2017.
4. *Forum Kamdar, Anubhav Malhotra, Pritish Mahadik*, "Display Message on Notice Board using GSM," Research India Publication, no. 7, vol. 3, pp. 827-832, March. 2016.

Secure Smart Office Automation System

Poornima Patil

Student, PCE, New Panvel

poornimakipat16de@student.mes.ac.in

Ashutosh Mohanty

Student, PCE, New Panvel

mohashu15@student.mes.ac.in

Anand Deshmukh

Student,PCE,New Panvel

anan15e@student.mes.ac.in

Saumitra Kulkarni

Student,PCE,New Panvel

mks15e@student.mes.ac.in

Krishnendu Nair

Faculty,PCE,New Panvel

knair@mes.ac.in

Abstract— The main idea behind the project is to design a smart system to provide security in offices/workplaces. The system performs several functionalities. One of them is providing authorized access to office via biometric so that only valid person can enter the office. The system also acts as an intruder detection system i.e. if any unauthorized person enters in the office, the owner as well as some desired people like watchman are informed about the activity. The system is said to be smart because it automatically turns all the lights and fans off whenever the last person in the office exits the office, thus conserving energy. The system is also designed to sense the presence of natural light in the room and manipulate the intensity of bulb accordingly. It also senses the current room temperature and sets the fan on/off accordingly. Adding on to all the mentioned functionalities the system starts the alarm if there is fire in the office. In conclusion it can be said that the system is designed to provide security as well as comfort at work places.

Keywords— Biometric, Sensors, IOT, Microprocessor

1. Introduction

As rapid change in technology always aims to serve the mankind, the expectation for living a simple yet advance and safe life keeps on increasing. Now a days office environment security is a major requirement of every individual when away from home or at the home. Office environment should be leisurely so that the employees can give their best as office environment directly affects the working efficiency of employees/workers. A smart office is a place that makes life easy for employees, which empowers it and increases their ability to stay connected[2]. A smart office aims to create a safe environment for employees so that they can focus more on their work and worry less about the safety. Sometimes

employees need to maintain confidentiality about some sensitive documents, because of this one has to always make sure whether he/she locked the door properly or not, his/her laptop is safe or not and many other things. A smart office is a system that does all this work for you with some more additional features that ensures comfort of the employee while working. The systems also contributes in conserving the energy making it efficient to use.

Internet of things(IOT) forms the base of the smart office system. The internet of things is a computing concept that describes the idea of everyday physical objects being connected to the internet and being able to identify themselves to other devices. The Internet of things (Iot) devices not only controls but also monitors the electronic, electrical and various mechanical systems which are used in various types of infrastructures. These devices which are connected to the cloud server are controlled by a single user (also known as admin) which are again transmitted or notified to all the authorized user connected to that network. Various electronics and electrical devices are connected and controlled remotely through different network infrastructures. Web browser present in laptop or mobile phone or any other smart technique through which we can operate switches, simply removes the hassle of manually operating a switch. Now a day's although smart switches are available they proves to be very costly, also for their working we required additional devices such as hub or switch .As there is rapid change in wireless technology several connectivity devices are available in the market which solves the purpose of communicating medium with the device and the microcontroller. Starting from Bluetooth to WiFi, from ZigBee to Z-wave and NFC

all solve the purpose of communicating medium[1]. In this project we have used Arduino UNO to control various devices.

2. Literature Survey

Office automation is a challenging one not only to the developer but also to the consumer. Developer has to choose the component as per the customer requirement. Due to all the customer demands are not equal hence they have to compromise with the existing products.

• *Lalit Mohan Satapathy, Samir Kumar Bastia and Nihar Mohanty (2018)*

Proposed a paper in which system is server independent and uses Internet of things to control human desired appliances starting from industrial machine to consumer goods. The user can also use different devices for controlling by the help of web-browser, smart phone or IR remote module. This paper presents a low cost flexible and reliable home automation system with additional security using Arduino microcontroller, with IP connectivity through local Wifi for accessing and controlling devices by authorized user remotely using Smart phone application.[1]

• *Renuka Bhuyar and Saniya Ansari (2016)*

Proposed a paper in which system is based on subsystems like lighting ,heating. Security and alarming systems are also present. The sensors are used to extract the real time data from environment. Sensors are connected to the ARM11 Controller. It processes the data and gives the output. Fan, bulb, buzzer are output devices connected to the controller which will work when the system crosses the threshold value. The sensor's data is continuously recorded. Fingerprint Identification module is used for security purpose. Fire alarm and emergency call is given to the service room. This data is stored in PC. This data

can be viewed on other PC's through Network switch. The data can be seen on the web page and on GUI.[2]

• *Neha Gabal, Neelam Barak and Shipra Aggarwal (2016)*

Proposed a paper in which an advanced approach to motion detection for automatic video analysis has been presented. The proposed method is a pixel dependent and non-parameterized approach that is based on first frame to build the model. The detection of the foreground which represents the object and background which is the surrounding of the environment starts once the subsequent frame is captured. It utilizes unique tracking methodology that identifies and eliminates the ghost object from dissolving into the background of the frame.[3]

• *Balakrishna Gokaraju, Donald Yessick, Jonathan Steel, Daniel A. Doss and Anish C. Turlapaty (2015)*

Proposed a paper in which intrusion detection system will be integrated wirelessly to the home WiFi system and could initiate an email to the respective authority. Moreover, these systems have high false alarm rates and unnecessary calls to 911 operator. The novelty of our present implementation design lies in cost and time effective communication of the intrusion event wirelessly to the home owners and law-enforcement with a confirmed image of the scene during the intrusion event[4].

2.1 Summary of Related Work

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

Table 1 Summary of literature survey

Sr. no	Paper	Conclusion/Summary
1.	Lalit Mohan Satapathy, Samir Kumar Bastia, Nihar Mohanty,2018, "Arduino based home automation using Internet of things (IoT)	The experimental setup which we designed has its focal point on controlling different home appliances providing 100% efficiency.
2.	Renuka Bhuyar Saniya Ansari, 2016, "Design and Implementation of Smart Office Automation System"	Many security safety techniques are used like smoke detectors ,illuminating and lighting
3.	Neha Gabal, Neelam Barak and Shipra Aggarwal,2016 "Motion Detection, Tracking and Classification for Automated Video Surveillance"	The results of the technique presented in paper have been analysed under qualitative and quantitative point of view. The results proved the efficiency of method on scales of accuracy and low processing requirements.
4.	Balakrishna Gokaraju, Donald Yessick, Jonathan Steel, Daniel A. Doss and Anish C. Turlapaty,2015"Integration of intrusion detection and web service alarm for home automation system using 'arm' microprocessor	The performance of the total integrated system was analyzed and tested over multiple iterations and found to be very robust in reliability of the signal strength and latency of web service alarm.

3. Proposed Work

IOT plays an important role in home automation system. The use of IOT in offices to ensure security is the key aspect of smart office automation system. The system includes various sensors that sense the environment and detect any malicious activities. The sensor data is then processed and according to the input, that is the sensed data the system produces output.

3.1 System Architecture

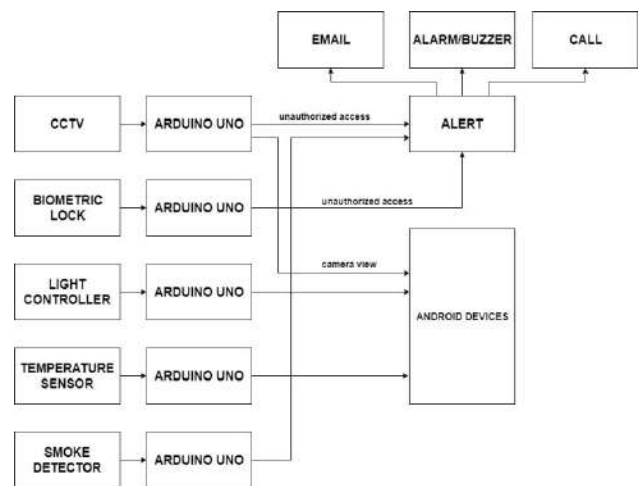


Fig. 1 Block diagram of smart secure office automation system.

- CCTV** : CCTV will be continuously capturing video footage. This video footage will be further processed using image processing techniques. It is used to capture activities of intruders if any.
- Biometric Lock**: Biometric lock will be implemented on the door, so that only authorized people can enter in the office.
- Light Controller**: Light controller continuously senses the indoor light intensity and accordingly controls the intensity of bulb present in the room. This helps in conserving energy.
- Temperature Sensor**: This sensor senses the room temperature and accordingly sets the fan on/off.
- Smoke Detector**: Smoke detectors are used to alert

people as soon as possible in case of fire. When smoke is detected in the room the alarm is set.

6. **Arduino UNO:** All the sensors are connected to separate arduino uno. The data from the sensors is taken as input and accordingly output is produced. If cctv detects any intruder activities the same data is sent to arduino which in turn sets the alarm, sends desired people email/call. The same is done if biometric lock comes across any unauthorized access and also when smoke detector detects smoke in the room.
7. **Email, Call:** Email and/or call will be sent to people who are registered in emergency contacts.
8. **Alarm:** In case of fire, any unauthorized access sensed by fingerprint lock, or any intruder activities are discovered the alarm situated in the office is set on.

4. Requirement Analysis

The setup is installed on a computer that includes some hardware and software components.

3.1 Software

The computer must have a dual core 64-bit processor with Windows operating system. The system must be installed with Android SDK(Software Development Kit) as well as Arduino IDE(Integrated Development Environment).

3.2 Hardware

As shown in Fig. 1 we set the devices then connect them according to the block diagram. All the sensors are connected to a separate Arduino uno. A camera will be installed at the all the red alert/sensitive places including the door. The arduino connected to camera will be in turn connected to android devices and alarm. Fingerprinting sensor is installed right outside the door which is connected to a solenoid lock that will lock/unlock the door. The system include humidity and temperature sensor - RHT03, a PIR sensor and a smoke detector which will be installed on the ceiling inside the office. A 12 volt battery is used as backup if there is power failure. Arduino uno kit is required for arduino setup. A alarm/buzzer to alert if there is any malicious activity

happening or even in case of fire to urgently evacuate the office premises.

5. Applications

This new wave of connectivity is going beyond laptops and smartphones, it's going towards smart homes, smart cities and military bases. Basically a connected life. These devices will bridge the gap between physical and digital world to improve the quality and productivity of life, society and industries. With IoT catching up Smart homes is the most awaited feature, with brands already getting into the competition with smart appliances. With the help of IoT security can be maintained in office and house.

1.Smart Home

Smart Home clearly stands out, ranking as highest Internet of Things application on all measured channels. More than 60,000 people currently search for the term "Smart Home" each month. This is not a surprise. The IoT Analytics company database for Smart Home includes 256 companies and startups. More companies are active in smart home than any other application in the field of IoT. The total amount of funding for Smart Home startups currently exceeds \$2.5bn. This list includes prominent startup names such as Nest or AlertMe as well as a number of multinational corporations like Philips, Haier, or Belkin.

2. Smart City

Smart city spans a wide variety of use cases, from traffic management to water distribution, to waste management, urban security and environmental monitoring. Its popularity is fueled by the fact that many Smart City solutions promise to alleviate real pains of people living in cities these days. IoT solutions in the area of Smart City solve traffic congestion problems, reduce noise and pollution and help make cities safer.

3.Military-Smart Bases

Incorporating IoT devices and sensors into military bases can have several positive effects. Automated security screening, for example, increases safety while decreasing manpower, and a network of security cameras connected to their environment via sensors and to a central network via the Internet will also minimize security risks. Smart management of resources – electricity and water for example – will increase the capacity and output of

military bases while ensuring that the wellbeing of all individuals inside the base is protected.

ACKNOWLEDGMENT

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

We would like to express deepest appreciation towards Dr. Sandeep joshi, Principal PCE, New Panvel, Dr. Madhumita Chatterjee, Head, Department of Computer Engineering Department whose invaluable guidance supported us in completing this project.

At last we must express our sincere heartfelt gratitude to all the members of Computer Engineering Department who helped us directly or indirectly during this course of work.

We would like to say that it has indeed been a fulfilling experience working on this project.

REFERENCES

1. [1]Lalit Mohan Satapathy, Samir Kumar Bastia, Nihar Mohanty, "Arduino based home automation using Internet of things (IoT)", 2018
2. [2]Renuka Bhuyar Saniya Ansari, "Design and Implementation of Smart Office Automation System",2016
3. [3]Neha Gabal, Neelam Barak and Shipra Aggarwal, "Motion Detection,Tracking Classification for Automated Video Surveillance",2016
4. [4]Balakrishna Gokaraju, Donald Yessick, Jonathan Steel, Daniel A. Doss and Anish C. Turlapaty, "Integration of intrusion detection and web service Alarm for home automation system using 'arm' microprocessor",2015
5. [5]Xin Hong ,Chenhui Yang,Chunming Rong Year, "Smart Home Security Monitor System",2016
6. [6]Zhen He,Ning Sun,Xiao liang,Yongchun Fang, "Wireless communication based smoke detection system",2016
7. [7]Jayasree Baidya, Trina Saha, Ryad Moyashir, Rajesh Palit, "Design and Implementation of a Fingerprint Based Lock System for Shared Access", 2015
8. [8]Ying-Wen Bai and Yi-Te Ku, "Automatic Room Light Intensity Detection and Control Using a Microprocessor and Light Sensors",2008
9. [9]Prof.S.A.ShaikhI and Aparna S.Kapare, "Intelligent Office Area Monitoring and Control Using Internet of Things",2017
10. [10]Sahana H S, Sandeep V S, Shwetha R, Sowmya J, Krupa K S, "Office Automation System Using Internet of Things",2017
11. [11]Prof. P. R. Rodge, Jaykant Prajapati, Anup Salve, Pallavi Sangle, "IoT Based Smart Interactive Office Automation",2017
12. [12]W. Li, T. Logenthiran, W. L. Woo, "Intelligent Multi-Agent System for Smart Home Energy Management",2015
13. [13]Adiono, T., Harimurti, S., Manangkalangi, B. A., & Adijarto, W., "Design of smart home mobile application with high security and automatic features" 2018
14. [14]Khan, A., Al-Zahrani, A., Al-Harbi, S., Al-Nashri, S., & Khan, I. A., " Design of an IoT smart home system" 2018
15. [15] ShariqSuhail, M., ViswanathaReddy, G., Rambabu, G., DharmaSavarni, C. V. R., & Mittal, V. K., " Multi-functional secured smart home" 2016

Prediction of Movie Box-office success using NLP and Machine Learning Techniques

Shreya Jayachandran, Sanika Tamhankar, Karishma Netake, Sayoojya Dinesan , and Dr. Sharvari Govilkar

Department of Computer Engineering, Pillai College of Engineering, Navi Mumbai, India - 410206

Abstract— Predicting movie success has always been a topic of great concern for producers and directors since its outcome is non deterministic outcome. Since these entities invest in movies, they need some kind of assurance that the movie will be successful in terms of financial goals. The proposed idea is to predict whether the movie will financially achieve its goal beforehand its completion. This will help investors associated with this business for avoiding investment risks. The prediction engine can be implemented using Natural Language processing and different machine learning and deep learning techniques like SVM and Multilayer Perceptron . There is a huge amount of data related to movies available over the internet from sources such as IMDb, Box Office Mojo, Rotten tomatoes etc. that can be used as an input to build the proposed model. Prediction can be done based on several important pre-released or post-released features like budget, audience voting, number of screens . The aim is to predict the success of a movie with maximum accuracy.

Keywords— Machine learning, Support vector machine, Neural Networks, Naive Bayes, K-nearest Neighbors, Logistic Regression, Decision Trees.

1. Introduction

With the emergence of Big Data, the world has changed in ways previously unimaginable. A lot of things which used to be impossible to analyze and predict even a decade ago, have now become easier and more intuitive to predict. The entertainment industry is no different. Over the years, data analytics and its immense power has been the root cause of a paradigm shift in the way operations are executed in the entertainment industry. In a rapidly growing and thriving industry such as the motion picture industry, data analytics has opened a number of important new avenues that can be used to analyze past data, make creative and marketing decisions, and accurately predict the fortunes of impending movie releases. The model can help investors to avoid risks and make a right choice of investment. Early prediction will help an investor to make choice if he/she wants to invest for new artists. It will also help producers to analyze conditions for success of movie and create insights for movie, maximizing the financial success.

2. Literature Survey

A. Predicting Movie Box Office Profitability: Travis Ginmu Rhee's paper discusses a classification model that are produced as a result of manual feature engineering using existing ones. Before that, the data is cleaned and normalized. Support Vector Machine which works well on binary classification problems is used to build the model. Evaluation of model is done using accuracy, confusion matrix and ROC-AUC. One more approach which is defined that of a neural network model whose performance is evaluated using cross entropy. One key observation is that both neural network and SVM approach have difficulty in classifying flop movies which can be overcome by including more movies in the dataset.

B. Early prediction using NLP techniques: A system is developed in which movie data and film scripts are used as a base for building the model. Scripts are processed using the NLTK package in python followed by Latent Semantic Analysis which are then vectorized using TF-IDF. The scripts are separated by genre to result in movie comparison based on scripts within the specific genre. Decision Tree, Random Forest, Naive Bayes and Support Vector Machine are the classification algorithm used and 10-fold Cross Validation is applied while training the models. Model is evaluated using accuracy, precision and recall as the evaluation metrics.

C. Relationships between Social Factors and Box Office Collections: In a paper proposed by Vinay Biramane, predictive models are built by establishing links between classical features, social media features and the overall success of the movie which includes total box office collection and the critics rating or review. The results show that the prediction model built using integration of classical as well as social factors can achieve higher accuracy rate. The predictive model is built using the machine learning techniques. The final product is a web application for which backend is designed in R language with support of shiny framework. For hosting the

application, services of shiny server, which supports R language, was utilized.

D. Prediction of Movies' popularity: Muhammad Hassan Latif proposed a technique to predict the popularity of a movie using linear regression, SVM regression and logistics regression. Movies targeted are the ones that were released from year 2004 to 2014 in order to get the latest trend of Market. WEKA which is used here is well suited for data mining tasks with collection of machine learning algorithms. It can perform classification, data preprocessing, clustering, regression, visualization and association rules.

E. Prediction of Movie Success for Real World Movie Data Sets: The model aims was to categorize the movie as successful or unsuccessful using given features in the data. Ratings of multiple users is fed as the input for the movie, and the outcome is a generalized and accurate rating that depicts the user's view of the movie. The methodology used is Fuzzy logic. Fuzzy string matching is used to determine an expressive linguistic review for a numerically calculated rating within the range of 0 to 10.

F. Predicting Movie Success: Cary D. Butler proposed a way to predict how successful a movie will be prior to its arrival at the box office. A total of five machine learning algorithms K-nearest Neighbor, Gaussian Naïve Bayes, Decision Trees, K-means Clustering, and Graphing Theory were applied to a dataset comprised of movie data from two different sources (IMDB and YouTube). Movie trailer data was collected in the form of views, likes, dislikes, and comments from YouTube. After applying the models and comparing K-means and graph theory, it was found that the graph theory model performs better than k-means or other three models.

G. Using Consumer-Centric Models: Using the potential of an unreleased movie, the model aims to predict movie success. Movie metadata is collected from various sources like IMDb, Box Office Mojo etc. For an unreleased movie, it is first shown to a test audience and then their ratings are appended to the data. Additionally, the audience are requested to rate other released movies which are also appended to the existing dataset. This is a consumer-centric approach that involves gathering user's taste and behaviour. It is observed that inclusion of this user rating data reduces the error by a factor of 2x. The results showed that the Linear regression depends heavily on good feature selection. Using all the features resulted

in bad performance. Hence some manual feature engineering will be required.

H. A Machine Learning Approach to Predict Movie Box-Office Success: Nahid Quader presents a simple approach by using SVM and a Multilayer Perceptron to build the model. The difference observed here is the use of both pre-released and post-released features. Text is processed using Text Analytics API and sentiment analysis is performed on them. A 10-fold cross validation is employed while training the model with SVM algorithm. Multilayer Perceptron gives a better prediction accuracy. Both the SVM and MLP approaches give good results but MLP produces a better prediction accuracy. Though MLP stands out, it can play much better with more data in hand and can be further tested with different architectures.

2.1 Summary of Related Work

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

Table 1: Summary of literature survey

Year of Publication	Research Paper details	Observation/Remarks
2016	Predicting Movie Box Office Profitability - A Neural Network Approach (IEEE).	SVM and a 3-layered back propagation neural network is used which was validated using cross entropy. BPNN was accurate than SVM with an accuracy of 91%.
2016	Early prediction of a film's box office success using natural language processing techniques and machine learning by Sean O'Driscoll.	Decision Tree, Random Forest Naive Bayes, SVM are the algorithms used. Word to vector conversion is done using TF-IDF. Metrics Used: Accuracy, precision, recall.
2016	Relationships between Classical actors, Social Factors	Models are built by establishing links between classical features, social media

	and Box Office Collections.	features and the overall success of the movie.
2016	Prediction of Movies popularity Using Machine learning Techniques.	Algorithms used include linear regression, SVM regression and logistic regression. Neural network is also used but logistic regression gives a better prediction accuracy of about 85%. Data mining is done using WEKA.
2017	Prediction of Movie Success for Real World Movie Data Sets.	Fuzzy string matching used to determine an expressive linguistic review for a numerically calculated rating within the range of 0 to 10. The accuracy obtained using fuzzy logic is 85%.
2017	Predicting Movie Success using Machine learning Algorithms.	Algorithms used include Logistic regression, Multilayer Perceptron, J48, Naive Bayes and PART. Results tested against k-fold cross validation. Best results are achieved through simple logistic regression around 84%.
2017	Improving Box Office Result Predictions for Movies using Consumer-Centric Models.	Algorithms used are KNN, Gaussian Naïve Bayes, Decision Trees, K-means Clustering, and Graphing Theory. KNN outperforms here with minimum error.
2018	A Machine Learning Approach to Predict Movie Box-Office Success (IEEE).	Performed sentiment analysis using text analytics APIs. NN gives a better prediction accuracy of around 89% than SVM due to the architecture used and its ability to extract important features.

3. Proposed Work

The input to the system would be the dataset taken from IMDB bollywood movies and we increased the dataset by adding more records. The first unit is Data cleaning. This will be followed by Pre-processing and Feature Engineering. Machine Learning techniques will then be applied on the dataset. The programming language 'python' is used for the implementation. The next part is comparing the results of different algorithm and finally the best algorithm will be chosen to predict the hit/flop result of the movie.

3.1 System Architecture

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

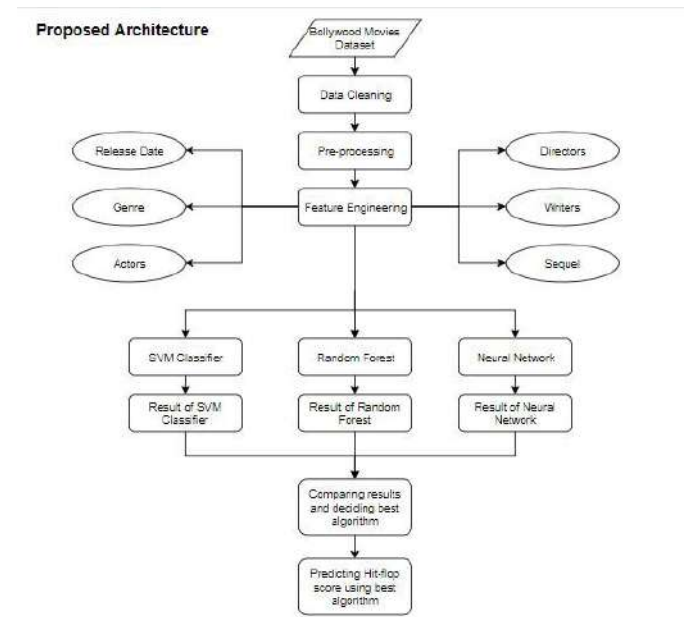


Fig. 1 Proposed system architecture

A. Bollywood Movies Dataset: The data is extracted from the IMdB site with movies released in the years between 1990 and 2018.

B. Data Cleaning: The second part receives the input data in a structured manner and then performs

cleaning operations on it in several steps. Data cleaning is the process of detecting and correcting corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the redundant data. Data cleaning is done by the Pandas and NumPy libraries in python.

C. Pre-processing: After the data is cleaned, the next step would be to preprocess it using certain methods. Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues.

D. Feature Engineering: After pre-processing is done, features are transformed or engineered. Using domain knowledge of the data features are created that make machine learning algorithms work.

E. Model building using machine learning algorithms: Once the features are ready to be given as input to the machine, model can be built or trained. The process of training an ML model involves providing an ML algorithm (that is, the learning algorithm) with training data to learn from. The learning algorithm finds patterns in the training data that map the input data attributes to the target (the answer that we want to predict), and it outputs an ML model that captures these patterns.

E. Result evaluation: Once the model is built, the one which gives best results can be considered for further predictions. The results obtained after training and testing the models using machine learning algorithms and neural networks are compared and evaluated to come to a final conclusion regarding choice of winning model.

F. Prediction: After choosing the best model, the final part of the working is to use the winning model to predict the hit/flop results of the test cases.

3. Requirement Analysis

The implementation detail is given in this section.

3.1 Software

Table 2: Software details

Operating System	Windows 10
------------------	------------

Programming Language	Python
Data storage	MS Excel

3.2 Hardware

The experiment setup is carried out on a computer system which has the hardware specifications as given :

Table 3: Hardware details

Processor	2.3 GHz Intel
HDD	1 TB
RAM	8 GB

3.3 Dataset and Parameters

The dataset comprises of movie entries that are extracted from the IMDb site. It comprises of features like release date, actors, writers, directors, genre that can be engineered to form useful parameters for the machine learning model. The engineered features can include actor-director pair, release month, etc to improve the prediction accuracy.

ACKNOWLEDGMENT

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

REFERENCES

1. Travis Ginmu Rhee, Farhana Zulkernine, "Predicting Movie Box Office Profitability - A Neural Network Approach (IEEE), Dec. 2016"
2. Sean O'Driscoll, "Early prediction of a film's box office success using natural language processing techniques and machine learning", Dec. 2016.
3. Vinay Biramane, Himanshu Kulkarni,

- “Relationships between Classical Factors, Social Factors and Box Office Collections (IEEE)”, Jan. 2016.
4. Muhammad Hassan Latif, Hammad Afzal, “Prediction of Movies popularity Using Machine Learning Techniques”, vol. 16, pp. 127-131, Aug. 2016.
 5. Sanjai Pramod, Abhisht Joshi, Geetha Mary, “Prediction of Movie Success for Real World Movie Data Sets”, vol. 3, pp. 445-461, Dec. 2017.
 6. Cary D. Butler, Eric Jackson, “Predicting Movie Success using Machine Learning Algorithms”, 2017.
 7. Rui Paulo Ruhrlander, Matthias Uflacker, “Improving Box Office Result Predictions for Movies using Consumer-Centric Models”, pp. 655-664, Aug. 2017.
 8. Nahid Quader, Md. Osman Gani, “A Machine Learning Approach to Predict Movie Box-Office Success (IEEE)”, Dec 2017.

Generalized Sentiment Learner Using Deep Learning

Ketaki Barde, Shweta Achary, Inderjeet Singh, Methu Manoharan, and Mentor: Prof Varunakshi Bhojane

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

Abstract

Newspapers and blogs express opinion of news entities (people, places, things) while reporting on recent events. We present a system that assigns scores indicating positive or negative opinion to each distinct entity in the text corpus. Our system consists of a sentiment identification phase, which associates expressed opinions with each relevant entity, and a sentiment aggregation and scoring phase, which scores each entity relative to others in the same class. Finally, we evaluate the significance of our scoring techniques over large corpus of news and blogs and any general text document.

Although, Semantic word spaces have been very useful, they cannot express the meaning of longer phrases in a principled way. Further progress towards understanding compositionality in tasks such as sentiment detection requires richer training, evaluation resources and more powerful models of composition.. To address challenges like sarcasm, slangs, abbreviations , we use the **Recursive Neural Tensor Network(RNTN)**. This model outperforms all methods of basic NN models on several metrics. It pushes the state of the art in single sentence positive/negative classification and the expected accuracy is about 80% up to 85.4%.

Introduction

News can be good or bad, but it is seldom neutral. Although full comprehension of natural language text remains well beyond the power of machines, the statistical analysis of relatively simple sentiment cues can provide a surprisingly meaningful sense of how the latest news impacts important entities. Sentiment Analysis is used for this soul reason. Opinion Mining is largely applied to data that comes with self-labeled information such as movie reviews on imdb. A scalar score comes along with the review text a user writes, which provides a good and reliable labelling of the text polarity. This ability to identify the positive or negative sentiment behind a piece of text is even more interesting when it comes to social data. Twitter gets new user data literally every second. If our model can predict sentiment labels for incoming live tweets, we'd be able to understand the most recent user attitude towards a variety of topics from a commercial flight satisfaction to brand image. The speciality of our system is that it exhibits the ability to process any general text document regardless of being structured or unstructured.

This system uses a logistic regression baseline model and complex-structured neural networks, Recursive Neural Network(RNN) and Recursive Neural Tensor Network(RNTN). Considering the nature of randomized data, it is first preprocessed, and a binary dependence tree is built which is fed as the input to the RNTN. Hyper-parameters need to be tuned and regularization methods such as L2 regularization are used as dropouts to optimize the performance. Features such as ‘Sentiment Index Formulation ‘

help to construct a statistical index which meaningfully reflects

the significance of sentiment term juxtaposition. This technique of using juxtaposition of sentiment terms and entities along with frequency weighted interpolation happiness levels aids to score entity sentiment. Next, statistical evaluation of the validity of our sentiment by correlating our index with several classes of real-world events is performed. Stanford Sentiment Treebank and a powerful Recursive Neural Tensor Network accurately predict the compositional semantic effects present in any text corpus.

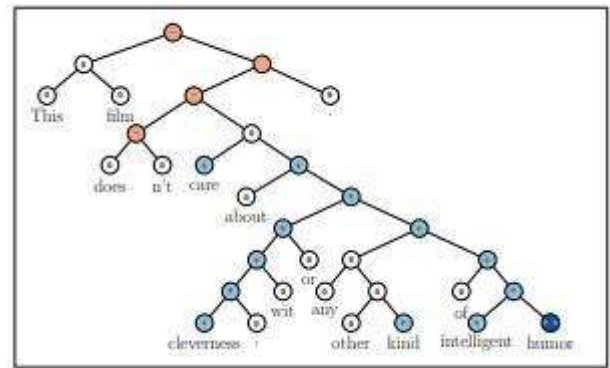


Figure 1: Example of the Recursive Neural Tensor Network accurately predicting 5 sentiment classes, very negative to very positive (– –, –, 0, +, ++), at every node of a parse tree and capturing the negation and its scope in this sentence.

Literature Review

1. Sentiment Analysis in Twitter using Machine Learning

A dataset is created by taking 600 positive tweets and 600 negative tweets. Feature extraction is needed due to presence of emoticons, slang words, misspellings, etc. It is done in two phases. In the first phase, twitter specific features are extracted. Then these features are removed from the tweets to create normal text. After that, again feature extraction is done to get more features. This is the idea used in this paper to generate an efficient feature vector for analyzing twitter sentiment. After feature extraction, sentiment classification is done. Three types of basic classifiers (SVM, Naive Bayes, Maximum Entropy) and ensemble classifier are used for sentiment classification. They obtained an accuracy of 90% whereas Naive Bayes has 89.5%.

2. Twitter Sentiment Analysis with Recursive Neural Networks

This paper shows how we experiment with different genres of Neural Net and analyze how models suit the data set using one hidden layer RNN, 2 hidden layer RNN and Recursive Neural Tensor Net (RNTN). Also different data filtering layers, such

as ReLU, tanh, and drop-out also yields many insights wrt performance. Tweets(limit 140) consists of emoticons, abbreviations, slangs or long tailing(happyyyy) hence normalization and data cleaning is needed before data set digraphs are fed as input. Data set used is (SemEval-2013 ,York University,6092 rows) having 5 levels-negative, objective, neutral, objective-OR-neutral and positive.Evaluation Metric used is accuracy and F1 average score. Balance of the data set and available labels of intermediate levels play a significant roles in training such models. Tuning the hyperparameters, regularization help to obtain a decent performance.

3. Learning Word Vectors for Sentiment Analysis

The strategy to learn word vectors specifically for the task of sentiment analysis is used. An unsupervised model is used to learn the semantic similarities between words, and a supervised component that is able to capture nuanced sentimental information. Given a document's bag of words vector v , features are obtained from the model using a matrix-vector product Rv , where v can have arbitrary tf weighting. For finding polarity v is not cosine normalized, instead cosine normalization is applied to the final feature vector Rv . Dataset used-IMDB dataset with 50000 reviews(approx. Equal no. of +ve and -ve reviews).

4. Applying Recurrent Neural Networks to Sentiment Analysis of Spanish Tweets

The dataset for evaluation considers annotated tweets with 4 polarity labels (P,N,NEU,NONE),positive negative neutral and NONE for absence of sentiment polarity. A Recurrent neural network architecture composed of Long Short Term Memory(LSTM) followed by feedback network is proposed. RNN composed of LSTM cells parse the input into a fixed-size vector representation which is used to perform the sentiment classification. Two variations of this architecture are used: (i) LSTM that iterates over the input word vectors OR (ii) over a combination of the input word vectors and polarity values from a sentiment lexicon.The general architecture of the model takes as inputs the words vectors and the lexicon values for each word from an input tweet. The inputs are then passed through a one-layer LSTM with a tunable number of hidden units. The generated representation is then used to determine the polarity of the input text using a feedforward layer with softmax activation as output function. The output of this last layer encodes the probability that the input text belongs to each class.

5. Large-Scale Sentiment Analysis for News and Blogs

This paper uses Sentiment lexicon generation to convolute semantic analysis with sentiment analysis. The concept of clustering is used to discover synonymous and anonymous words followed by computing the respective polarities. A statistical model is developed to classify groups adjectives into clusters, corresponding to their tone/orientation and dimensions such as general, health, crime, sports, business, politics, media and then polarity of each word(WordNet is used for classification) is determined which results in greater accuracy.

6. Twitter as a Corpus for Sentiment Analysis and Opinion Mining

Dataset- From Twitter API, magazines, popular newspapers,etc. The distribution of word frequencies in the corpus is checked and then TreeTagger is used to tag all the posts in the corpus. Steps followed:- 1]Filtering: Remove all

URLs and twitter account names. 2]Tokenization: segment text to form bag of words. 3]Removing stop-words(articles) from bag of words. 4]Constructing n-grams: make a set of n-grams out of consecutive words. A negation is attached to a word which precedes it or follow it (eg. a sentence "I do not like fish" will form two bigrams: "I do+not", "do+not like";"not+like fish"). They have trained two classifiers, which use different features: presence of n-grams and part-of-speech(POS) distribution information. N-gram based classifier uses the presence of an n-gram in the post as a binary feature. POS based distribution estimates probability of POS-tags presence within different sets of texts and uses it to calculate posterior probability.

7. A study of Sentiment Analysis using Deep Learning Techniques on Thai Twitter Data.

In this paper, two deep learning techniques are used : Long Short Term Memory (LSTM) and Dynamic Convolutional Neural Network (DCNN). Word2vec is used to train initial word vectors for LSTM and DCNN models.LSTM has 3 internal gates: input gate, forget gate and output gate. DCNN has various layers. The dynamic k max pooling layer performs a selection of the top maximum k values on the column of sentence matrix. The dynamic k value is used instead of fixed k value. This pooling strategy makes DCNN suitable for any various lengths of input. Each network has arbitrary number of convolutional layer, folding layer, and pooling layer connected together. In the labeling process, pre-classified emotions are used to classify tweets into positive or negative. Tweets are labelled to each class when it contains emoticons corresponding to only that class. 3-folds cross validation is used in the verification process. Both techniques give significantly higher accuracies than classical techniques such as NB and SVM, but not MaxEnt.

8.Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank

Uses a new model called the Recursive Neural Tensor Network (RNTN) to capture the compositional effects with higher accuracy (80.7%). Dataset used is Stanford Sentiment Treebank. This model take as input phrases of any length and represent a phrase via word vectors & a parse tree thereby computing vectors for higher nodes in the tree using the same tensor-based composition function. Lastly, test set is used comprising of positive and negative sentences and their respective negations to show that, unlike bag of words models, the RNTN accurately captures the sentiment change and scope of negation thereby learning that sentiment of phrases following the contrastive conjunction 'but' dominates.Features such as Matrix based Compositionality algorithms, Recursive Auto-associative memories capture similarities and evaluate sentiment polarity. Recursive neural models compute parent vectors in a bottom up fashion using different types of compositionality functions. The parent vectors are again given as features to a classifier yielding greater accuracy.

9.Rationalizing Sentiment Analysis in Tensorflow

The paper describes a two-part model which predicts a multi-sentiment analysis (called encoder) and extracts summary phrases (called generator). The encoder (enc) is a supervised learning problem which predicts a rating given a text review. Training samples are (x, y) pairs, where $x = \{x_t\}_{T=1}^T$ is an input text sequence of length T and $y \in [0, 1]^m$ is an output vector where m denotes the number of aspects in a review.The generator (gen) is a text summarization task which

selects a subset of words of the text review as a rationale describing rating. There is no target rationale, but rather both the encoder and generator are jointly trained. The output of the generator are probabilities of each word in the original review being selected as part of the rationale. That is, the final predictions are trained on the rationale output from the generator, and not the full text review.

10. Sentiment Analysis on Movie Reviews using Recursive and Recurrent Neural Network Architectures

Several approaches associated with RNN model were discussed such as Computing Semantic Word Vector, Affine NN, based on Mean Likelihood, etc. The most prominent one was Recursive-Recurrent Neural Network Architecture (accuracy 83.88%). In this approach, the movie reviews are split into sentences using tokenizer. Then, each sentence is fed into a Recursive Neural Network (RNN) which outputs a hidden vector and a sentiment for the sentence. The RNN's hidden vector of each sentence is passed as an input to a Recurrent Neural Network. This enables us to capture phrase-level sentiment for each sentence and sentence-level sentiments for the whole document. The IMDB movie review dataset is used. We train the word vectors on this corpus using the skip-gram architecture. We used this trained model on the classification task on the IMDB movie review dataset.

Summary Of Related Work

S N	Paper	Advantages and Disadvantages	Year
1.	Neethu M.S, Rajasree R	Advantage: Feature extraction is done multiple times which can increase the accuracy. Disadvantage: Takes considerably large amount of time.	2013
2.	Ye Yuan, You Zhou	Advantage: Experimenting with different genres of neural net and choosing the best one. Disadvantage: Emoticons are removed.	2016
3.	Andrew L. Maas, Raymond E. Daly, et.al	Advantage: Use of an unsupervised model to learn the semantic similarities between words, and a supervised component that is able to capture nuanced sentimental information. Disadvantage: Can not process large amount of data.	2011
4.	Oscar Araque, Rodrigo Barbado, et.al	Advantage: Two variations of RNN {LSTM cells} used. Disadvantage : Lack of better pre-processing at word level.	2017
5.	Manjunath Srinivasaiah, Namrata Godbole, et.al	Advantage: Developing statistical model to classify adjectives and then finding polarity which increases accuracy. Disadvantage: Doesn't show how sentiment varies by demographic group, news	2007

S N	Paper	Advantages and Disadvantages	Year
6.	Alexander Pak, Patrick Paroubek	Advantage: POS based distribution estimates probability of POS-tags presence within different sets of texts and uses it to calculate posterior probability. Disadvantage: Checking the distribution of words frequencies in the corpus and then using TreeTagger to tag all the posts in the corpus, all of which has to be done before data cleaning step.	2010
7.	Peerapon Vateekul, Thanabhat Koomsubha	Advantage: The use of dynamic k pool layer makes DCNN suitable for any lengths of input. Disadvantage: Due to 3 fold cross validation, accuracy increases but time to process increases as well.	2016
8.	Richard Socher, Alex Perelygin, et.al	Advantage: Use of Sentiment Treebank for sentiment compositionality. Parent vectors are given to classifier as features yielding better accuracy. Disadvantage: Currently not suitable for multilingual analysis, cautious pre-processing and regularization required.	2013
9.	Alyson Kane, Henry Neeb, et.al	Advantage: Prescribes a two-part model which predicts a multi-sentiment analysis (called encoder) and extracts summary phrases (called generator). Disadvantage: Comparatively less precision.	2017
10.	Aditya Timmaraju, Vikesh Khanna	Advantage: Capturing phrase-level sentiment for each sentence and sentence-level sentiment for the whole document. Disadvantage: Emoticons are removed.	2015

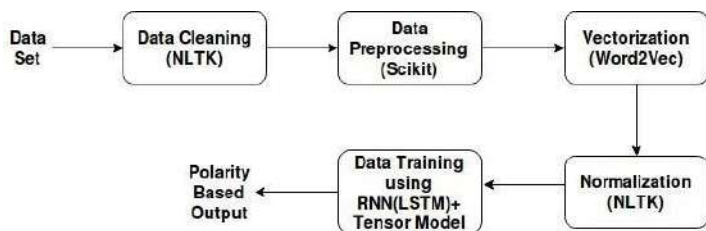
3. Proposed Work

In order to capture the compositional effects with higher accuracy, a new model called the **Recursive Neural Tensor Network (RNTN)** can be used. Recursive Neural Tensor Networks take as input phrases of any length. They represent a phrase through word vectors and a parse tree and then compute vectors for higher nodes in the tree using the same tensor-based composition function. One can use recursive neural tensor networks for boundary segmentation, to determine which word groups are positive and which are negative. The same applies to sentences as a whole. Word vectors are used as features and serve as the basis of sequential classification. They are then grouped into sub

phrases, and the sub phrases are combined into a sentence that can be classified by sentiment and other metrics.

3.1 System Architecture

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



3.1.1 Data Set :

The dataset will constitute of online available Twitter bank, Stanford Sentiment Data resources like the Treebank and IMDB data sets. Also, online reviews, news articles, blog texts, book/movie/restaurant reviews and any text format readables can be processed. The data is collected in the CSV file. The collected data has to be partitioned into 80% for training and 20% for testing. Various aspects of supervised learning shall be used to train the model. The Stanford Sentiment Treebank is the first corpus with fully labeled parse trees that allows for a complete analysis of the compositional effects of sentiment in language. The corpus is based on the dataset introduced by Pang and Lee (2005) and consists of 11,855 single sentences extracted from movie reviews. It was parsed with the Stanford parser (Klein and Manning, 2003) and includes a total of 215,154 unique phrases from those parse trees, each annotated by 3 human judges. This new dataset allows us to analyze the intricacies of sentiment and to capture complex linguistic phenomena.

3.1.2 Data Cleaning and Preprocessing:

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Pre-processing includes removal of emoticons, slangs, lengthy abbreviations(eg.happyyyy). A recursive neural network requires the training data to have a predetermined tree structure. A PCFG Stanford NLP Parser[3] can be used to build estimates of the actual optimal tree structures. One can run the parser basing on a careless probabilistic context-free grammar model, which works better than traditional PCFG models on less strictly grammatical input data such as tweets in our case. Moreover, the recursive neural network assumes each non-leaf node to have two children. So binarization of parse tree is mandatory using a binarizer based on Michael Collin’s English head finder. After these processes, all non-leaf nodes in our parse tree have at most two children. It is possible that a node has only one child, for example NP → N. We chose to soft delete this node in our NN implementation where cost and errors are directly passed to the next level without modification at this level.

Neural networks are much more powerful than our baseline logistic regression model because they can learn complex intermediate units(neurons) and capture nonlinear interactions between inputs. They are also prone to overfitting for the same reason. They are so powerful that they usually fit noises in the training data as well as the general model. In order to generalize the model to unseen data sets, we put a lot of

emphasis on regularization method like dropout and L2 standard..

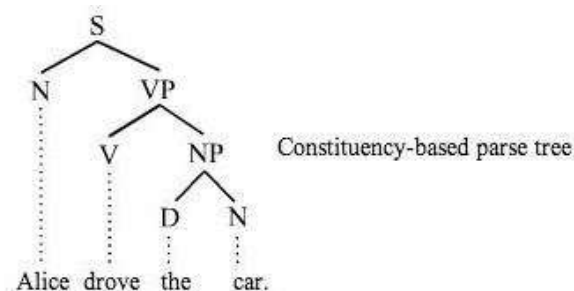
Python libraries-NLTK/Scikit aid in this task. It has several methods which help in Scaling, Binarization, Case Conversion, trimming etc.

3.1.3 Data Vectorization and Normalization

Word vectors are used as features and serve as the basis of sequential classification. They are then grouped into sub phrases, and the sub phrases are combined into a sentence that can be classified by sentiment and other metrics. Recursive neural tensor networks require external components like Word2vec, which is described below. To analyze text with neural nets, words can be represented as continuous vectors of parameters. The first step toward building a working RNTN is word vectorization, which can be accomplished with an algorithm known as Word2vec. Word2Vec converts a corpus of words into vectors, which can then be thrown into a vector space to measure the cosine distance between them; i.e. their similarity or lack of. Word2vec is a separate pipeline from NLP. It creates a lookup table that will supply word vectors once you are processing sentences.

Normalization refers to scaling and formatting to meet similar standards. Meanwhile, your natural-language-processing pipeline will ingest sentences, tokenize them, and tag the tokens as parts of speech. To organize sentences, recursive neural tensor networks use constituency parsing, which groups words into larger sub phrases within the sentence; e.g. the noun phrase (NP) and the verb phrase (VP). This process relies on machine learning, and allows for additional linguistic observations to be made about those words and phrases. By parsing the sentences, you are structuring them as trees.

The trees are later binarized, which makes the math more convenient. Binarizing a tree means making sure each parent node has two child leaves.

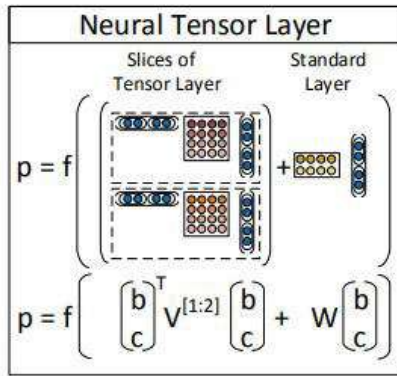


The entire sentence is at the root of the tree (at the top); each individual word is a leaf (at the bottom). Finally, word vectors can be taken from Word2vec and substituted for the words in the generated tree.

3.1.4 Data Training and Operation :

In the standard RNN, the input vectors only implicitly interact through the nonlinearity (squashing) function. A more direct, possibly multiplicative, interaction would allow the model to have greater interactions between the input vectors. Therefore a new model RNTN is used wherein the main idea is to use the same, tensor-based composition function for all nodes.

Principal Dr. Sandeep Joshi who provided us the opportunity to do research on this domain and present this work. Furthermore, we would like to acknowledge with much appreciation the crucial role of Prof. Rupali Nihare and Prof. Gaurav Sharma for their guidance in selecting this project and also for providing us all the details for proper presentation of this project. We extend our sincere appreciation to all our Professors from Pillai College of Engineering for their valuable insight and tips throughout the work.



Above figure depicts a single layer of the Recursive Neural Tensor Network. Each dashed box represents one of d -many slices and can capture a type of influence a child can have on its parent.

The RNTN uses this definition for computing p_1 :

$$p_1 = f \left(\left(\begin{bmatrix} b \\ c \end{bmatrix}^T V^{[1:d]} \begin{bmatrix} b \\ c \end{bmatrix} + W \begin{bmatrix} b \\ c \end{bmatrix} \right) \right)$$

where W is as defined in the previous models. The next parent vector p_2 in the tri-gram will be computed with the same weights:

$$p_2 = f \left(\left(\begin{bmatrix} a \\ p_1 \end{bmatrix}^T V^{[1:d]} \begin{bmatrix} a \\ p_1 \end{bmatrix} + W \begin{bmatrix} a \\ p_1 \end{bmatrix} \right) \right)$$

The main advantage over the previous RNN model, which is a special case of the RNTN when V is set to 0, is that the tensor can directly relate input vectors. Intuitively, we can interpret each slice of the tensor as capturing a specific type of composition

Hence using advanced Python libraries to extrapolate patterns and exploiting sigmoid functions to output polarity levels, a significant increase in output is expected.

Requirement Analysis

The experiment setup is carried out on a computer with different hardware and software requirements as specified below.

1. Hardware

In order to run this project successfully, we require an Intel Processor of minimum 2GHz, Hard Disk Drive(HDD) of 180GB and a RAM of minimum 4GB.

2. Software

RNN is supported by Python programming language. Also several other tools such as SciKit, NLTK which are used in the processing phase are in-built libraries provided by Python. Therefore, we will use Python and Windows Operating System which provides the platform for execution of the Python program. It is suitable for any OS, Python being platform independent.

Acknowledgement

We take this opportunity to express our gratitude and deep regards to our Guide Prof. Varunakshi Bhojane for the guidance, monitoring and constant encouragement throughout the course of this thesis. We would also like to thank our Head

Conclusion :

The main advantage of our system is that it exhibits the ability to process any generalized textual document. The high accuracy RNTN model exploits several features to produce polarities as output. Currently, we have worked towards producing accurate results for any English scripted documents. However, future scope may entail processing capabilities of Hinglish script, emoticons, sarcasm as well as slangs. This system can be applied to study customer behavioral patterns, market study, lifestyle reviewing (Books/Movies/Restaurants) and opinion mining. Sentiment analysis strikes for cautious pre-processing and the proper model that best fits the data set. Balance of the data set and available labels of intermediate levels play a significant role in training such models. Imbalance of our data set leads to a poor performance and an under-fit in RNTN. Another take-home lesson would be tuning the hyperparameters for a better data fit.

References:

- [1] Bernard J. Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. 2009. Micro-blogging as online word of mouth branding. In CHI EA '09: Proceedings of the 27th international conference extended abstracts on Human factors in computing systems, pages 3859–3864, New York, NY, USA. ACM. [url : https://nlp.stanford.edu/~socherr/EMNLP2013_RNTN.pdf](https://nlp.stanford.edu/~socherr/EMNLP2013_RNTN.pdf)
- [2] A. Merin. 1999. Information, relevance, and social decision making: Some principles and results of decision theoretic semantics. In Lawrence S. Moss, Jonathan Ginzburg, and Maarten de Rijke, editors, Logic, Language, and Information, volume 2. CSLI, Stanford, CA. [url : https://cs224d.stanford.edu/reports/YuanYe.pdf](https://cs224d.stanford.edu/reports/YuanYe.pdf)
- [3] Twitter GloVe word vectors. <http://nlp.stanford.edu/projects/glove/> Retrieved on May 6th 2015.
- [4] Neethu M.S, Rajasree R, "Sentiment Analysis in Twitter using Machine Learning", IEEE, 2013
Reference - <https://ieeexplore.ieee.org/document/6726818/>
- [5] Peerapon Vateekul, Thanabhat Koomsubha, "A study of sentiment analysis using deep learning techniques on Thai Twitter data", IEEE, 2016.

Reference
<https://ieeexplore.ieee.org/document/7748849/>

[6] Ye Yuan, You Zhou, "Twitter Sentiment Analysis with Recursive Neural Networks", 2016

Reference -

<https://cs224d.stanford.edu/reports/YuanYe.pdf>

[7] Manjunath Srinivasaiah, Namrata Godbole, et.al, "Large-Scale Sentiment Analysis for News and Blogs", 2007

Reference -

<http://www.uvm.edu/pdodds/files/papers/others/2007/godbole2007a.pdf>

[8] Richard Socher, Alex Perelygin, Jean Y. Wu, "Recursive Deep Models for Semantic Comp", 2013

Reference -

https://nlp.stanford.edu/~socherr/EMNLP2013_RNTN.pdf

[9] Aditya Timmaraju, Vikesh Khanna, "Sentiment Analysis on Movie Reviews using Recursive and Recurrent Neural Network Architectures", 2015

Reference -

<https://cs224d.stanford.edu/reports/TimmarajuAditya.pdf>

[10] Andrew L. Maas, Raymond E. Daly, "Learning Word Vectors for Sentiment Analysis", 2011

Reference - <http://www.aclweb.org/anthology/P11-1015>

[11] Alexander Pak, Patrick Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", 2010

Reference -

<http://crowdsourcing-class.org/assignments/downloads/pak-paroubek.pdf>

[12] Oscar Araque, Rodrigo Barbado, J. Fernando S'anchez-Rada y Carlos A. Iglesias, "Applying Recurrent Neural Networks to Sentiment Analysis of Spanish Tweets", 2017

Reference -

http://ceur-ws.org/Vol-1896/p8_gsi_tass2017.pdf

[13] Alyson Kane, Henry Neeb, et.al, "Rationalizing Sentiment Analysis in Tensorflow", 2017

Reference -

<https://web.stanford.edu/class/cs224n/reports/2758389.pdf>

POST WI-FI CHAT ANDROID APP

MANUSCRIPT TRACK: SOFTWARE TECHNOLOGIES SYSTEMS FOR FUTURE CITIES

Sonal Jotiba Adsol(PCE,Student), Vaibhavi Deepak Dalvi (PCE,Student), Srikanth Bashyam Naidu (PCE,Student), Aditya Balakrishnan Warriar(PCE,Student), Gayatri Hegde(PCE,Faculty).

Abstract:

Along with the rapid growth of Wi-Fi-enabled smartphone devices, the need for efficient content sharing techniques for smartphones in Wi-Fi environment has increased. There are many methods recently presented by various researchers for efficient content sharing in Smartphones However existing methods are suffered from the various limitations. Hence, this area of research is still thought provoking problem for researchers. The paper present secure approach for message passing and content sharing in smartphones using Wi-Fi (WCS), which gives better speed and range over Bluetooth and no internet connection or heavy centralized servers, are required in this process. Each smartphone can work as client and server at the same time .Message passing is done securely by using encryption decryption technique. Users can also share files with large numbers of other users in the network following Peer-to-Peer(P2P) protocol. Based on the analysis done for secure message passing scheme, it is observed that with little difference in time secure message passing is done using encryption.Also We can See Our current location Via GPS in the designed app. When there is loss of coverage area Internet connection is lost then also people can communicate to each other through the app.

Keywords:

File sharing, Map tracking, P2P, Audio calling,Texting Encryption,Decryption .

Submitted on: 30 October 2018

Revised on:

Accepted on:

***Corresponding Author Email:**sonaljadsol@gmail.com

Phone:8108820175

I. INTRODUCTION

Smartphones have become very popular and commonplace in the last few years. They have become highly capable multimedia devices that can share various types of user generated contents and messages by making use of data plans over the internet. Messaging is another important feature and application in mobiles. Messaging is possible over the internet and off the internet (through SMS service).One traditional technique of sharing on mobile phones is done through MMS (Multimedia Messaging Service), but it has a size limitation. A sophisticated data exchange model Bluetooth standard that we use today is allowing wireless communication between enabled devices. However sending data using Bluetooth in the real world is not easy. Devices should follow the Bluetooth standard[3] to discover devices via Bluetooth and connect to them. According to some researchers, Bluetooth is not an appropriate solution to some situations

Following issues are there in P2P file sharing application for mobile devices in ad-hoc environments:

- ❑ How to organize the shareable contents for a mobile device?
- ❑ How a mobile device can discover the content of users interest in proximity?
- ❑ How efficient transmission of content from one/more mobile devices to one/more other mobile devices can be ensured?
- ❑ How energy consumption in mobile devices is minimized to save battery power?

Solution to above problem is peer-to-peer sharing among Smartphones, by using Wi-Fi instead of using expensive packet data networks for sharing files source text by using the linguistic method to interpret and examine the text. The abstractive summarization aims to produce a generalized summary, conveying information in a concise way.

LITERATURE SURVEY

Some mobile P2P file sharing applications like Mobile Mule, Symella, mbit have been developed [6][7].

For all these applications:

- (1) Devices need to be connected with internet and
- (2) It is assumed that created network is stable and infrastructure based. Another real time problem with these applications is internet connectivity is not always very reliable, and even if it is available, it incurs high usage cost. To avoid this cost, a mobile user can use the Wi-Fi interface of the mobile phone to connect to the Internet. Mobile P2P content sharing using ad-hoc environments is a popular area of study [1],[3],[6],[7]. These studies are promising, but they focus on routing technique and assume that the underlying network is multi-hop. In such case, smart- phones (peers) that are within reach of each other can form a mobile ad-hoc network using Wi-Fi (IEEE 802.11), and find efficient way to discover and download contents generated by other peers in the network. 802.11 supports an ad-hoc mode of operation, provides longer communication range as compared to Bluetooth, and now a day all smartphones are equipped with the Wi-Fi feature. This paper includes the proposal of a system for providing the above features of content sharing and messaging over the Wi-Fi. ANDROID platform is used for deployment of an application that provides secure messaging and content distribution using Wi-Fi without using the internet or costly centralized servers.

II. IMPLEMENTATION

i. FUNDAMENTAL

Smartphones have become very popular and common place in the last few years. They have become highly capable multimedia devices that can share various types of user generated contents and messages by making use of data plans over the internet. Messaging is possible over the internet and off the internet (through SMS service). One traditional technique of sharing on mobile phones is done through MMS (Multimedia Messaging Service), but it has a size limitation.

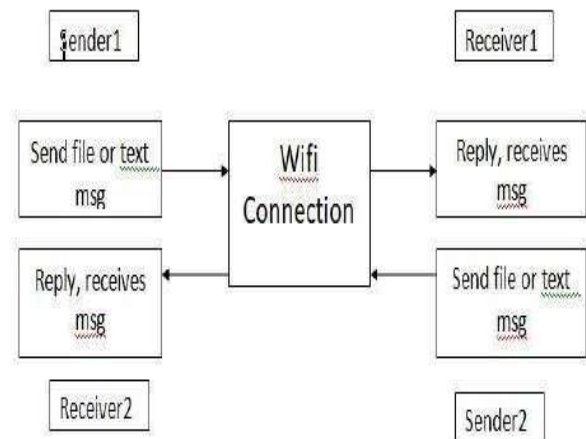


Fig 1. User connectivity

To make the communication easy and reliable we used the Wi-Fi connection. On both the end the sender and the receiver should enable the Wi-Fi connection of their smartphone. By enabling the Wi-Fi connectivity the sender and receiver can communicate easily they can text each other, share the files between them, etc.

III. SYSTEM ARCHITECTURE

The Goal is to implement the secure message passing and content sharing scheme for smart phones using wireless network. And to get current location via GPS. Messages are transmitted by sender and stored on server database in encrypted form. From server database they are delivered to receiver's Smartphone if he is online otherwise delivery of message is delayed till user is not online. Contents to be stored are kept in sharable folder with each smartphone user and. Metadata of all these files is kept with server. Actual file is not present with server. Any file from sharable folder can be downloaded by requester. In System we are having wireless Adhoc network, inside of that handsets are communicating with each other through peer to peer connection. Central server having database of smartphone users which have already done their registration. Central server has GUI, Request Response Manager and File Transaction Tracker. Android Mobile Handsets have Communication Manager, Shared File Folder and message processor.

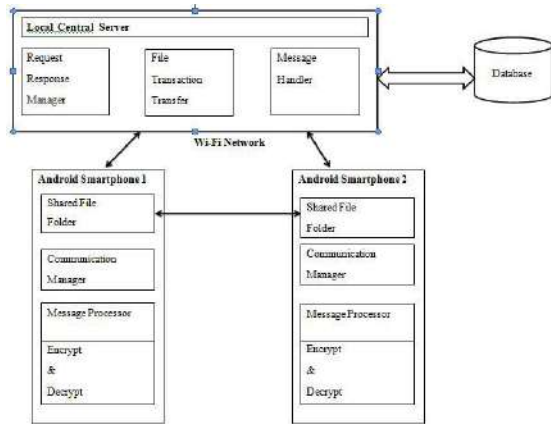


Fig 2. System Architecture

Client Side

1. Communication Manager: It is responsible for handling the connectivity of the system, both It deals with sending request to the server for searching contents and receiving list of peers containing required contents.
2. Shared File Folder: It allows user to make content available to other peers.
3. Message processor: At sender side it does the work of encrypting message and at receivers end it decrypts message.

Server Side

The server handles registering the users into the network, validating users and searching and showing the list of files available in the network to a user.

1. File Transaction Tracker: It keeps the track of transactions carried out by use
2. Request Response Manager: It handles request from user and provides response accordingly.
3. GUI : It provides interaction to administrator.
4. Message Handler: Stores and retrieves encrypted messages from database.

i. FILE AND MESSAGE SHARING TECHNIQUE

Each user has folder of sharable files .When user log into the system his IP address and list of shared files is updated on server. Any user can get the list of shared files with owner's IP address by sending request to server. For sharing message user asks

server for list of network users and then selecting User_id of peer user can share chat message with it. If user clicks on to User_id of any peer all previous communications with that user can be viewed. If user is offline then messages intended for him will get stored on server database in encrypted form and once he is offline all stored messages will get delivered to him and decrypted messages can be viewed at receivers end.

ii. ALGORITHM DESIGN

For secure message passing new encryption and decryption algorithm is design. Procedure followed for encrypting and decrypting message are briefly given here.

Encrypting Message :

1. Before sending string S to server do $Ascii(S[i])$ is processed as $E_k(S[i]) = \text{Encrypted } S[i]$ endfor
2. Provide this encrypted string (Es) to Huffman algorithm to encode E(S) in the form of 1 and 0 .
3. Resulting String H(Es) is stored in server database.

Decrypting Message:

1. H(Es) is processed with Huffman decoding technique to get Es.
2. Decrypt Es by using Dk to get S again. $D_k(Es) = S$
3. Resulting plain text S is displayed on users device

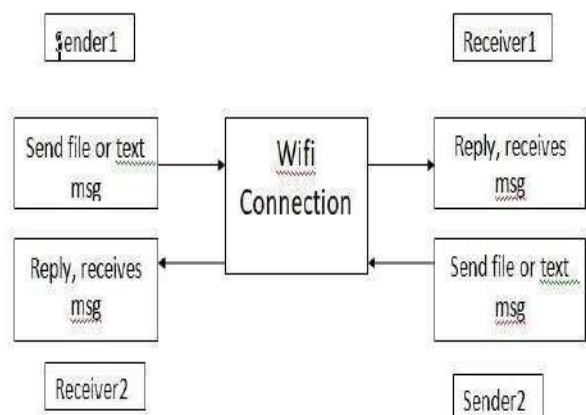


Fig3. User connectivity

iii. **FEATURES**

- No Internet connection : No internet connection is needed while communicating.
- Audio calling : User can make a call without using internet.
- File transferring : User can transfer files when they are connected to each other without using internet.
- Text message : The main feature of chat application is sending and receiving Text message without using internet.
- GPS Location : User can access map while travelling without using internet connection.

iv. **HARDWARE AND SOFTWARE REQUIREMENTS**

Hardware Requirements:

- 128MB RAM
- Pentium 3 Processor Speed 500MHz
- MIN 5GB HDD

Software Requirements:

- OS : WINDOWS 7,8,XP
- Language used : JAVA ,HTML JDK 1.8, XML

v. **ADVANTAGES**

- Doesn't require Internet connection.
- Different features would be provided like file sharing, audio calling.
- Easy to handle.

vi. **APPLICATIONS**

- To send Text message : Sending and receiving text message.
- Audio calling : Audio calling can also applicable for user.
- GPS Location : Location can be obtained.
- File sharing : files can be shared between the users.

CONCLUSION

So , now with this chat application , people can communicate with each other via Wi-Fi connection where the internet connectivity is not available or having battery consumption problem .The purpose of this project is to implement a peer-to-peer

connection between the mobile phone at no cost. The system will allow users to search for other individuals within WIFI range and to establish free peer to peer connection for transmitting messages , sharing files and making voice communication .

REFERENCES

- [1.] Ghassan Kbar, Wathiq Mansoor ,“Voice over IP Mobile Telephony Using WIFI P2P”,2010 Sixth International Conference on Wireless and Mobile Communications.
- [2.] Vishal S. Kasat1 , Rakesh Pandit2 ,”Implementation of Mobile To Mobile Calling through Wi-Fi”, International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064,Jan. 2015.
- [3.] Prof. Govind Wakure, Tanzeel Shaikh ,Kirti Karande, Ibrahim Shaikh, Hardik Vaghela, “Bluetooth Message Hopping Chat Application,” (Information Technology, Rajiv Gandhi Institute of Technology, Versova,Mumbai.), Sept. 2015.
- [4.] Prof. Manjitsing valvi, Jay P. Chauhan , Dinsha(2015). “Development of WiFi-Bluetooth Communication Protocol” , Apr. 2015 .
- [5.] mbit.tv, “Mobile file sharing,” available online at <http://mbit.tv/index.jsp>.Last accessed on June 8, 2010.
- [6.] "Specification of the Bluetooth system, part B: baseband specification," The Bluetooth SIG, I Dec 1999.
- [7.] Piotr K. Tysowski, Pengxiang Zhao, Kshirasagar Naik, “Silent broadcast: Experience of connectionless messaging Using Wi-Fi P2P,” in Proc.of the (ICIDT), 2012 8th International Conference on Information Science and Digital Content Technology , June 2012, pp. 239–242.

Author Biographical Statements



Sonal Adsol is student of Computer Science and Engineering from Pillai College of Engineering , New Panvel , Maharashtra .



Aditya Warriar is student of Computer Science and Engineering from Pillai College of Engineering , New Panvel , Maharashtra .



Vaibhavi Dalvi is student of Computer Science and Engineering from Pillai College of Engineering , New Panvel , Maharashtra .



Prof. Gayatri Hegde currently working as Assistant Professor in Pillai College of Engineering , New Panvel. She has completed BE (Basveshwar Engineering College , Bagalkot , Karnataka) , ME (Pillai College of Engineering , New Panvel) . Currently Pursuing PhD (Thadomal College of Engineering , Mumbai) .



Srikaanth Naidu is student of Computer Science and Engineering from Pillai College of Engineering , New Panvel , Maharashtra .

Human Resource Analytics (HRA) For Employees Using Deep Learning

Shivang Panchal, Jayadev Venkatesh, Manpreet Singh Channa, Siddhesh Deshpande, Prof. Varunakshi Bhojane,

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

Abstract—Employees are a valuable asset in any organization. But if they quit their jobs unexpectedly, it may incur huge costs to any organization. Using Machine Learning, we will build a model which will both predict and explain whether employees will leave their employer or not and the reasons why they may do so. The data comprises a wide range of topics which allow to explain employee's leave behaviour in relation with:

Organizational Factors (department). Employment Relational Factors (i.e tenure, the number of projects participated in; the average working hours per month; objective career development; salary). Job Related Factors (performance evaluation).

Keywords—Machine Learning ,Deep Learning.

1. Introduction

Employee retention refers to the various policies and practices which let the employees stick to an organization for a longer period of time. Every organization invests time and money to groom a new joiner, make him a corporate ready material and bring him at par with the existing employees. The organization is completely at loss when the employees leave their job once they are fully trained. Employee retention takes into account the various measures taken so that an individual stays in an organization for the maximum period of time.

Research says that most of the employees leave an organization out of frustration and constant friction with their superiors or other team members. In some cases low salary, lack of growth prospects and motivation compel an employee to look for a change. The management must try its level best to retain those employees who are really important for the system and are known to be effective contributors. It is the responsibility of the managers as well as the management to ensure that the employees are satisfied with their roles and responsibilities and the

job is offering them a new challenge and learning every day.

2. Literature Survey

In today's scenario where the world is trying to sustain the economic development for maintaining a steady growth on the business end where employees are being an asset and important resources to any organization is also a treat to the sustainability of the organization. We observed that the type of model that obtained results with the most accuracy is inconsistent. The best performing model throughout the survey varies. In [1], Random Forest outperformed other models. Similarly, [2] demonstrates that XG-Boost gives the most accurate result. [3] demonstrates that SVM is the superior algorithm. Hence, the results obtained through the system proved to be dependent upon the nature of the provided datasets.

2.1 Summary of Related Work

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

Table 1 Summary of literature survey

Sr. No	Paper	Advantage
1.	Dilip Singh Sisodia, Somdutta Vishwakarma, Abinash Pujahari (2016)	Individual decision trees can be trained in parallel. Overall variance decreases and the number of base models increase.
2.	Rohit Punnoose, (2016)	Gradient Boosting can be used to optimize any differentiable loss function .
3.	Sepideh Hassankhani Dolatabadi, Farshid Keynia (2017)	SVM can model non-linear decision boundaries, and there are many kernels to choose from. They are also fairly robust against overfitting, especially in high-dimensional space.

3.	Sepideh Hassankhani Dolatabadi, Farshid Keynia (2017)	The time required for training grows quadratically with respect to the number of datapoints.
----	---	--

Table 2 Summary of literature survey

Sr.No	Paper	Disadvantage
1.	Dilip Singh Sisodia, Somdutta Vishwakarma, Abinash Pujahari (2016)	They're not easily interpretable. Large number of base models lead to more time in both training and testing.
2.	Rohit Punnoose, (2016)	Training generally takes longer because of the fact that trees are built sequentially.

3. Proposed Work

3.1 System Architecture

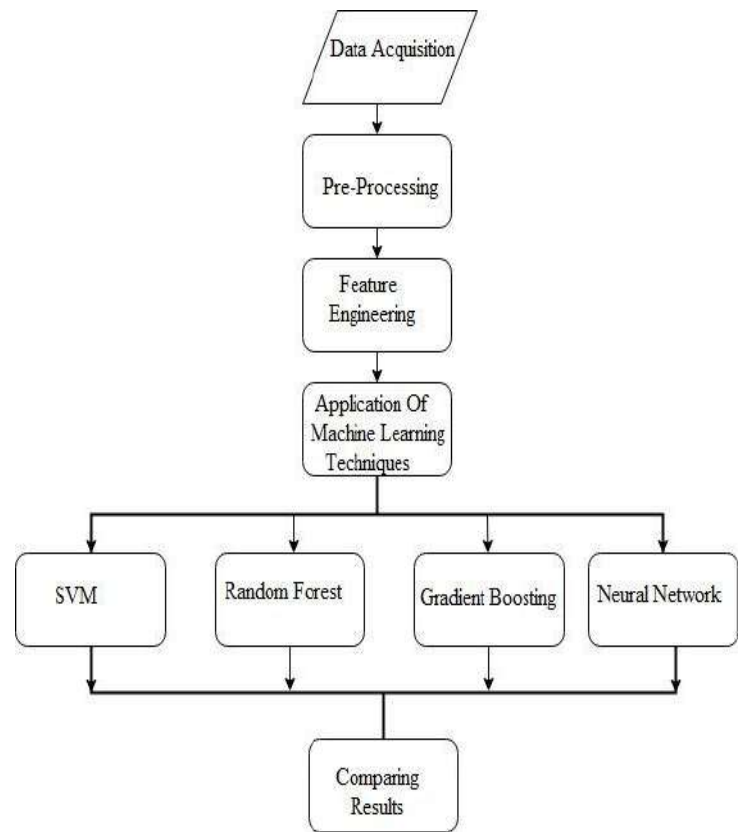


Fig. 3.1.1 Proposed System Architecture

Pre-Processing:

Handling missing values:

Different ways to handle missing values are :

1.) Delete the rows which contain missing values:

This should only be preferred if the number of rows having missing values is very less as compared to total number of rows.

2.) Imputation:

This process is also called as filling. Numerical features are filled with global mean, mode or median

of that feature and categorical features are filled with most frequent value.

3.)Imputation based on class:

Let's say we have a binary classification problem. Instead of filling with a global mean we can fill with the mean of the class to which the datapoint belongs to.

Feature Engineering:

Handling Categorical Features:

Any type of variable must be converted into a numerical variable. Categorical variables are of two types:

1.)Ordinal:

Ordinal features have an inherent order among them. One way to handle a categorical variable is to give a number to each category.

For example: In the feature Performance Rating, Low / Good / Excellent / Outstanding changes to 1 / 2 / 3 / 4

Since the numbers have an inherent order this technique is good for ordinal features but not for nominal.

2.)Nominal:

Nominal features are handled with the help of One-Hot encoding.

Application Of Machine Learning Techniques:

1.)Support Vector Machine:

SVM [4] is a supervised learning algorithm which can be used to perform classification as well as regression.It is based on the idea of finding a hyperplane that best separates the two classes.We try to find a hyperplane such that its distance from the nearest datapoint from both classes is maximum.The data points on either side of the hyperplane are called support vectors

2.)Random Forest:

Random forest [5] is an implementation of bagging ensemble method wherein all the base learners are decision trees.Each of the base learners will be a high variance model also called as strong learners.Even though all the base learners are high variance models the meta-model will have reduced variance.

3.)Gradient Boosting (XGBoost):

The idea behind additive modelling is to compute 'm' simple functions and then combine/add them to form

our final complex function.Gradient Boosting [6] uses additive modelling to gradually nudge/tweak an approximate model towards a really good model by adding simple sub models into a composite model.In boosting simple models are called weak learners.

4.)Neural Network:

An artificial neural network [8] is a biologically inspired computational model that is patterned after the network of neurons present in the human brain. Artificial neural networks can also be thought of as learning algorithms that model the input-output relationship. Applications of artificial neural networks include pattern recognition and forecasting in fields such as medicine, business, pure sciences, data mining, telecommunications, and operations managements.

Comparing Results:

Based on the data set we will predict whether the employee will quit the organization or not. The predicted value will be compared with the actual value in the dataset by using various evaluation metrics like precision, recall, confusion matrix, ROC curve.

3.Requirement Analysis

The implementation detail is given in this section.

3.1 Software

Table 3.1.1 Software details

Operating System	Windows 10
Programming Language	Python
Libraries	Pandas, Numpy, Sklearn, Matplotlib, keras

3.2 Hardware

Table 3.2.1 Hardware details

Processor	2.2 GHz Intel i7
GPU	4GB or more
RAM	8GB or more

3.3 Dataset and Parameters

Table 3.3.1 Sample Dataset Used

AGE	Attrition	Business Travel	Daily Rate	Department
41	Yes	Travel_Rarely	1102	Sales
49	No	Travel_Frequently	279	R&D
37	Yes	Travel_Rarely	1373	R&D

List of other features :

- Distance From Home
- Education
- Education Field
- Employee Count
- Employee Number
- Hourly Rate
- Job Involvement
- Job Level
- Job Role
- Job Satisfaction
- Marital Status
- Monthly Income
- Monthly Rate
- Number of Companies Worked
- Over 18
- Percent Salary Hike
- Performance Rating
- Relationship Satisfaction
- Standard Hours
- Stock Option Level
- Total Working Years
- Training Times Last Year

- Work Life Balance
- Years at Company
- Years in Current Role
- Years Since Last Promotion
- Environment Satisfaction
- Gender
- Years with Current Manager

CONCLUSION

In this paper, the study of different domain techniques is presented. The different techniques such as SVM, Random Forest, Gradient boosting and Neural Network are explained with examples. The different standard datasets or variable inputs are defined that may be used in experiment for this domain systems. Through Literature survey, we observed that the type of model that obtained results with the most accuracy is inconsistent. The best performing model throughout the survey varies. Hence, the results obtained through the system proved to be dependent upon the nature of the provided datasets. To determine the best performing model, comparative study of various techniques mentioned above is presented in this report. This can be calculated by implementing performance measures like precision, recall, confusion matrix and receiver operating characteristic (ROC) curve which are described in this report.

ACKNOWLEDGMENT

It is our privilege to express our sincerest regards to our supervisor Prof. Varunakshi Bhojane 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. Madhumita Chatterjee and our Principal Dr. Sandeep M. Joshi for encouraging and allowing us to presenting this work.

REFERENCES

- [1] Evaluation of Machine Learning Models for Employee Churn Prediction :Dilip Singh Sisodia, Somdutta Vishwakarma, Abinash Pujahari Department of Computer Science & Engineering National Institute of Technology Raipur, India (2016)

[2] Prediction of Employee Turnover in Organizations using Machine Learning Algorithms :A case for Extreme Gradient Boosting.Rohit Punnoose, PhD candidate XLRI – Xavier School of Management Jamshedpur, India (2016).

[3] Designing of Customer and Employee Churn Prediction Model Based on Data Mining Method and Neural Predictor,Sepideh Hassankhani Dolatabadi Department of Industrial Engineering Islamic Azad University of Technology Kerman, Iran (2017).

[4] C. Cortes and V. Vapnik, “Support vector machine,” *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[5] L. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, no. 5, pp. 1– 35, 1999.

[6] T. Chen and C. Guestrin, “XGBoost: Reliable Large-scale Tree Boosting System, 2015”, Retrieved from http://learningsys.org/papers/LearningSys_2015_paper_32.pdf. Accessed 12 December 2015.

[7]”HR Employee Attrition and Performance”, Available:

<https://www.ibm.com/communities/analytics/watson-analytics-blog/hr-employee-attrition/>

[8] Artificial Neural Network, <https://developer.nvidia.com/discover/artificial-neural-network>

[9] Existing System architecture referred from : <https://techinsight.com.vn/language/en/employee-churn-prediction/>

THORACIC DISEASES PREDICTION ALGORITHM FROM CHEST X-RAY IMAGES USING MACHINE LEARNING TECHNIQUES

Rushikesh Chavan

Student,PCE,New Panvel

rushibalucom17@student.mes.ac.in

Jidnasa Pillai

Student,PCE,New Panvel

jidnasavikum16de@student.mes.ac.in

Shravani Holkar

Student,PCE,New Panvel

shravaniudhol16de@student.mes.ac.in

Prajyot Salgaonkar

Student,PCE,New Panvel

prajyotps15it@student.mes.ac.in

Prof. Prakash Bhise

Faculty,PCE,New Panvel

pbhise@mes.ac.in

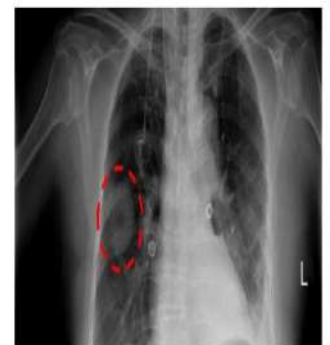
Abstract— Examining Chest X-Ray (CXR) is a time consuming process. In some cases, medical experts had overlooked the diseases in their first examinations on CXR, and when the images were reexamined, the disease signs could be detected. Radiologists have to spend time diagnosing these chest X-ray images to find any potential lung diseases. Diagnosing X-ray require careful observation and knowledge of anatomical principles, physiology, and pathology. In this work, we are applying traditional machine learning techniques for automated prediction of thoracic diseases from chest X-ray images. Computerized images segmentation and feature analysis can assist the doctors in treatment and diagnosis of diseases more accurately. In our approach, we are applying traditional machine learning techniques in building independent binary classifier for each of the diseases(Cardiomegaly, Edema, Emphysema, Hernia, Pneumonia, Fibrosis, Pneumothorax). Pre-processing of image gray scale image by resizing and cropping is done . Application of SIFT (Scale-invariant feature transform) computer vision algorithm on pre-processed image is to be done to detect feature descriptors in the image. Visual bag of words is constructed from feature descriptors obtained from the images. Computed visual bag of words is used as a feature vector for Logistic regression and SVM. Each model's output is binary label.

Keywords—Radiologists, chest X-ray, thoracic diseases, independent binary classifier, Scale-invariant feature transform,

Normal Lungs



Infected Lungs



Visual bag of words, Logistic regression, Support Vector Machine.

1. Introduction

Figure 1. Difference in normal chest x-ray and infected chest x-ray. [8]

Radiologists have to spend time diagnosing the chest X-ray images to find any potential lung diseases. Examining chest X-ray is one of the most frequent and cost effective medical imaging examination. Diagnosing x-rays require careful observation and knowledge of anatomical principles, physiology, and pathology. Developing automated system for such could make a huge impact to the patients, who don't have access to expert radiologists.

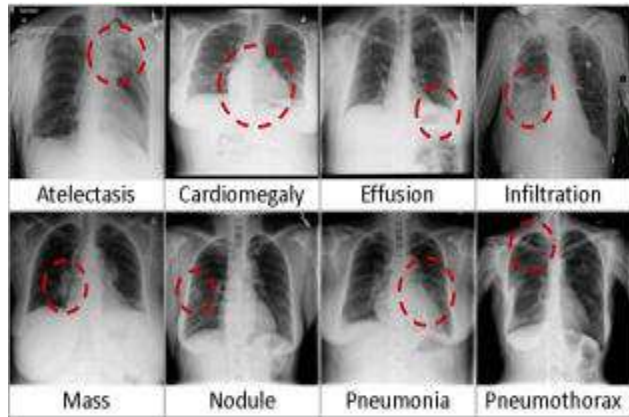


Figure 2. Eight common thoracic diseases observed in chest X-rays that validate a challenging task of fully-automated diagnosis.[7]

In our approach, we will be applying traditional machine learning techniques in building independent binary classifier for each of the diseases. We will preprocess gray scale image by resizing and cropping them. SIFT (Scale-invariant feature transform) a computer vision algorithm when applied on pre-processed image detects feature descriptors in the image. Visual bag of words will be constructed from feature descriptors obtained from the images. Computed visual bag of words is used as a feature vector for Logistic regression and SVM.

2. Literature Survey

1. *Chest X-RAY Analysis to detect Mass Tissues in Lungs :*

This work presents a method for abnormal mass tissue detection on digital x-ray. It adopted the template matching technique for detecting mass tissue. Although various research has done based on template matching for mass tissue detection,

this work adopted DCT based template matching which has decreased the matching time. This is suitable for real time x-ray image abnormality detection or detecting mass tissues from video image of x-ray. For simplicity this works has assumed that the mass tissue area will be 8x8, 16x16, 32x32, 64x64 in size. The mass tissue area size can be other

2. *Image Segmentation for Lung Region in Chest X-ray Images using Edge Detection and Morphology :*

Here [2] they have shared or experience on segmenting the lung shape on CXR image. The segmentation process starts by detecting the lung edge using canny edge detection filters. To improve the edge detection, Euler number method is applied. Later, morphology method is used to make the lung edge better so that the final output of lung region can be generated. After implementing the segmentation task, the output in the form of lung region mask is compared to the GT image to check their similarity. In the evaluation, the Jaccard Similarity Coefficient is used to calculate the similarity. The value derived from the test is moderately high although it cannot exceed other prior researchers score. In the future, the proposed segmentation method can be modified to be applied to other medical image types such the MRI and CT so that the ROI can be isolated from the other parts.

3. *Foreign Object Detection in Chest X-rays :*

The Authors (Zhiyun Xue, Serna Candemir, Sameer Antani, L. Rodney Long, Stefan Jaeger, Dina Demner-Fushman, George R. Thoma) focused on identifying one common type of foreign object shown in chest X-rays-buttons on gowns. The method consists of four major steps: image intensity normalization, low contrast image identification and enhancement, lung region segmentation, and button objects extraction. Two methods for the step of button objects extraction were applied. One is based on the circular Hough transform, the other is the Viola-Jones object detector.

4. Thorax Disease Diagnosis Using Deep Convolutional Neural Network :

They propose a new framework to augment the dataset dramatically. Using the augmented dataset to train a CNN model for the thorax disease diagnosis, they improve the model performance significantly. Their future work is to combine millions of images without labels collected from local hospital to improve the performance of the CNN models.

5. Automatic Detection of Major Lung Diseases Using Chest Radiographs and Classification by Feed - forward Artificial Neural Network :

Here the authors (Shubhangi Khobragade, Aditya Tiwari, c.Y. Pati and Vikram Narke) developed automated system for the detection of lung diseases such as TB; pneumonia and lung cancer using chest radiographs. From the results; we can say that image preprocessing techniques like histogram equalization; image segmentation gives good results for the chest radiographs. Pattern recognition technique such as feed forward artificial neural network is giving good results. The limitation of this proposed method is that it is not robust when there are changes in the size and position of chest x-ray image.

6. Detection of Pneumonia clouds in Chest X-ray using Image processing approach :

Computer assisted detection of diseases from CXR are always very helpful at places where there is shortage of skilled radiologist. In countries like India where, we do not have experienced radiologists in rural areas, such tools can be of immense help by automatically screening people who need urgent medical care and further diagnosis. The authors (Abhishek Sharma, Daniel Raju, Sutapa Ranjan) have identified the lung region by rib cage boundary identification. They have also used Otsu thresholding to segregate the pneumonia cloud from the healthy lung in the lung area, still working on other methods that can be adopted for thresholding the CXR images can yield better results. As pneumonia clouds are not visible in the image of the lung after Otsu thresholding,

this ratio is expected to be much lower than when computed for healthy lungs without clouds.

7. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases :

They attempted to build a “machine-human annotated” comprehensive chest X-ray database that presents the realistic clinical and methodological challenges of handling at least tens of thousands of patients (somewhat similar to “ImageNet” in natural images). They also conducted extensive quantitative performance benchmarking on eight common thoracic pathology classification and weakly-supervised localization using ChestX-ray8 database. The main goal is to initiate future efforts by promoting public datasets in this important domain. Building truly large-scale, fully-automated high precision medical diagnosis systems remains a strenuous task. ChestX-ray8 can enable the data-hungry deep neural network paradigms to create clinically meaningful applications, including common disease pattern mining, disease correlation analysis, automated radiological report generation, etc. For future work, Chest X-ray will be extended to cover more disease classes and integrated with other clinical information, e.g., follow up studies across time and patient history.

8. Effective Pneumothorax Detection for Chest X-Ray Images Using Local Binary Pattern and Support Vector Machine :

The primary method in this paper is to segment the lung in the abnormal region through multiple overlapping blocks. The abnormal region is found by texture transformed from computing multiple overlapping blocks. Finally, this method effectively analyses lung diseases of the area in the chest X-ray image and improves the possible diagnosis of the missing problem of the pneumothorax area. This increases the efficiency for physicians to assess the extent of the treatment of pneumothorax, so as to support the radiologist to reduce workload. This study presents a novel framework for automatic

pneumothorax detection in CXRs. The texture analysis is based on intensity and gradient for pneumothorax detection. The pneumothorax case was a difficult judgment when pneumothorax region is extremely stenotic and close to the chest boundaries. In addition, pixels located near the chest boundaries tend to have less discriminative texture on image indication, because the bones and pleura existed in obvious edges, which reduced their correspondence of textures. Consequently, the texture characteristic in chest boundaries area is not as prominent as in the inner lung region. Discrimination in different lung regions and adding the texture weight may be the future research focus. The segmentation can increase the accuracy rate for the segmentation of pneumothorax region.

2.1 Summary of Related Work

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

Table 1 Summary of literature survey

SR. No.	Literature	Techniques
1.	Chest X-RAY Analysis to detect Mass Tissues in Lungs	<ol style="list-style-type: none"> 1. Contrast-stretching 2. Histogram Normalization. 3. Discrete Cosine Transform Matching
2	Image Segmentation for Lung Region in Chest X-ray Images using Edge Detection and Morphology	<ol style="list-style-type: none"> 1. Edge detection using Canny Edge Detection and Euler Number Method. 2. Morphology Techniques (erode and dilate). 3. Jaccard Similarity Coefficient for calculating Similarity.
3.	Foreign Object Detection in Chest X-rays	<ol style="list-style-type: none"> 1. Image intensity normalization 2. Calculation of the mean value and standard deviation value of the image intensities for image enhancement. 3. Atlas based lung

		segmentation algorithm <ol style="list-style-type: none"> 4. SIFT flow Algorithm 5. Circle Hough Transform. 6. Viola-Jones algorithm for face detection. 7. Sobel and Canny Detector. 8. Edge gradient thresholding.
4.	Thorax Disease Diagnosis Using Deep Convolutional Neural Network	<ol style="list-style-type: none"> 1. Image alignment using Weber Local Descriptor. 2. Gaussian Scale Space Theory. 3. Data augmentation. 4. CNN architecture development based on the Caffe framework. 5. Component analysis using SST. 6. SIFT for classification.
5.	Automatic Detection of Major Lung Diseases Using Chest Radiographs and Classification by Feed - forward Artificial Neural Network	<ol style="list-style-type: none"> 1. Histogram Equalization. 2. Image Filtering using Histogram Equalization. 3. Thresholding. 4. Edge Detection. 5. Standard Deviation. 6. Entropy. 7. Artificial Neural Network. 8. Feed Forward Neural Network.
6.	Detection of Pneumonia clouds in Chest X-ray using Image processing approach	<ol style="list-style-type: none"> 1. Resizing. 2. Histogram Equalization. 3. Otsu Thresholding.
7.	ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on	<ol style="list-style-type: none"> 1. Natural Language Processing for detecting Pathology keyword. 2. DNorm a machine learning method for

	Weakly-Supervised Classification and Localization of Common Thorax Diseases	<p>disease recognition and normalization.</p> <ol style="list-style-type: none"> 3. MetaMap a prominent tool to detect bioconcepts from the biomedical text corpus. 4. Deep Convolutional Neural Network (DCNN) 5. Pre-trained models : ImageNet, e.g., AlexNet, GoogLeNet, VGGNet-16 and ResNet-50. 6. Ad-hoc thresholding based on B-Box generation method.
8	Effective Pneumothorax Detection for Chest X-Ray Images Using Local Binary Pattern and Support Vector Machine	<ol style="list-style-type: none"> 1. Uniform local binary pattern (ULBP). 2. Support vector machine- (SVM). 3. Otsu algorithm. 4. 8-connected neighborhood method. 5. Gaussian Filtering.

The overview of comparison of different parameters are given in Table 2

Table 2 .Summary of literature survey

SR No.	Literature	Dataset	Advantages	Disadvantages
1.	Chest X-RAY Analysis to detect Mass Tissues in Lungs	JSRT X-Ray Dataset.	This work adopted DCT based template matching which has decreased the matching time. This is suitable for real time	The size of mass tissues were assumed as 8x8,16x16,32x32 in size. But the mass tissues can be of any size.

			x-ray image abnormality detection or detecting mass tissues from video image of x-ray.	Which then can not be detected.
2.	Image Segmentation for Lung Region in Chest X-ray Images using Edge Detection and Morphology	JSRT X-Ray Dataset	Canny Edge Detection, Euler Number Method, Morphology Techniques, Jaccard Similarity gave result moderately high	The detection score does not exceed from the prior researches .
3.	Foreign Object Detection in Chest X-rays	NLM Indiana dataset	Requires only two methods to detect the foreign object	Identifying only one common type of foreign object shown in chest X-rays.
4.	Thorax Disease Diagnosis Using Deep Convolutional Neural Network	Dataset from Local Hospital	Due to the usage of augmented dataset there is an improvement in the model performance significantly	Limited images to train the CNN due to use of local hospitals datasets so the results are not that accurate.

5.	Automatic Detection of Major Lung Diseases Using Chest Radiographs and Classification by Feed - forward Artificial Neural Network	Sasoon Hospital; PUNE	Image preprocessing techniques like histogram equalization; image segmentation Pattern recognition technique such as feed forward artificial neural network is giving good results which is 92% accurate.	The limitation of this proposed method is that it is not robust when there are changes in the size and position of chest x-ray image.
6.	Detection of Pneumonia clouds in Chest X-ray using Image processing approach	JSRT X-Ray Dataset	Automated detection of pneumonia in short time, helpful especially in rural areas where there is a lack of skilled radiologists.	As pneumonia clouds are not visible in the image of lung after Otsu thresholding, ratio of detection is much lower than computed for healthy lungs without clouds.
7.	ChestX-ray8: Hospital-s	Chest X-ray 14 Dataset	ChestX-ray8 can enable the data-hungry	No use of Public Databases in order to

	cale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases		deep neural network paradigms to create clinically meaningful applications, including common disease pattern mining, disease correlation analysis, automated radiological report generation.	carry out the processing.
8.	Effective Pneumothorax Detection for Chest X-Ray Images Using Local Binary Pattern and Support Vector Machine	Dataset from Chung Shan Medical University Hospital, in Taichung, Taiwan.	This method effectively analyses lung diseases of the area in the chest X-ray image and improves the possible diagnosis of the missing problem of the pneumothorax area. This increases the efficiency for physicians to assess the extent of the treatment of pneumothorax, so as to support the radiologist to reduce workload.	Discrimination in different lung regions and adding the texture Weight is missing in this work, which can improve the accuracy of segmentation.

3. Proposed Work

Proposed work is focused mainly on using traditional feature extraction techniques like SIFT, machine learning algorithms like logistic regression, SVM and Computer Vision algorithm like Visual bag of words to produce prediction of lung diseases.

3.1 System Architecture

The system architecture is given in Figure 4

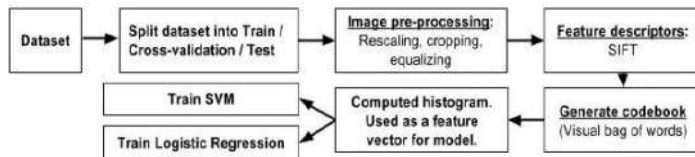


Figure. 3 Proposed system architecture

A. Dataset:

- Published by National Institutes of Health (NIH) Clinical Center
- 100,000+ frontal-view X-ray images
- 32,717 unique patients, 14 lung diseases
- Each image has multi-label
- Images are gray scale of size 1024 x 1024

B. Data split and pre-processing pipeline:

Data pipeline to split data, pre-process it, and for feature generation. Each image will be pre-processed by scaling from 1024 x 1024 to 224 x 224 to speed up computation. Rescaling followed by cropping to make lungs in the image focal, resulting in image of size 180 x 200. Image contrast will be increased by applying histogram equalizer. We will be splitting for train / Cross-validation / Test set, by randomly selecting ~20,000 unique patients for train, ~5000 for Cross-validation, and ~5000 for Test set. Since, each disease will be having independent binary classifier; separate dataset will be generated for each of the disease classifier. Images will be randomly sampled for randomly sampled patients. For each disease classifier, data will be balanced with equal number of sample for label-1

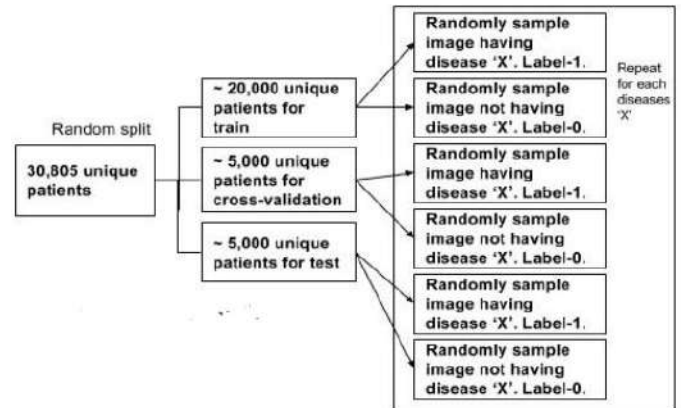


Figure 4 Data splitting for differentiating the lung diseases.

C. Feature Extraction:

For extracting features, we are applying SIFT to capture local information in the image. SIFT is a computer vision algorithm used to detect and describe local features in images. SIFT descriptor is invariant to translations, rotations and scaling transformations in the image and robust to moderate perspective transformations and illumination variations. SIFT first finds the key points within an image and then compute descriptor vector for each keypoint. Image is convolved with Gaussian filters at different scale, and then the difference of successive Gaussian blurred images is computed. Keypoints are the maxima/minima of the Difference of Gaussian (DoG) that occurs at multiple scales. Orientation is computed for each keypoints based on local image gradient directions. Using orientation, descriptor vector is computed for each keypoint.

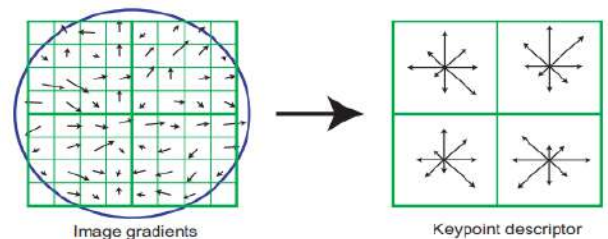


Figure 5. Descriptor Vector [16]

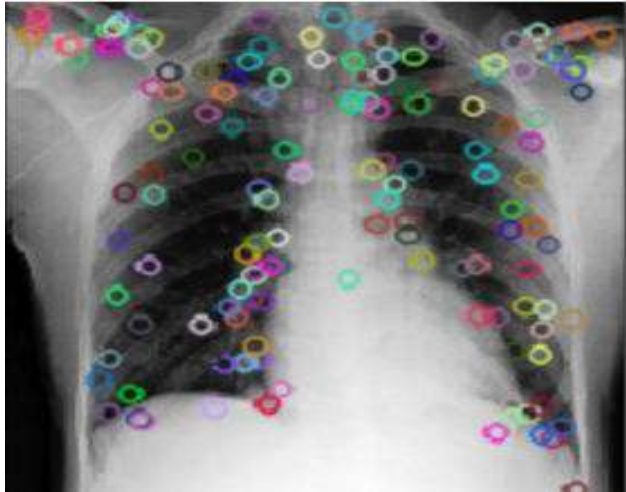


Figure 6. Keypoint Localization [15]

D. Codewords Dictionary:

Bag of Visual Words (Codebook) BoW model constructs a large vocabulary of visual words. For BoW, features are extracted using SIFT, then codebook will be generated, followed by histogram. K-means clustering is applied to extracted features from all image to generate codebook. Centroids are defined as a visual codewords. Size of the codebook is number of clusters. Each extracted feature is mapped to one of the closest centroid. Resulting histogram of for each image, which counts the number of features for each of the visual code words. Histogram is used as feature vector for training models.

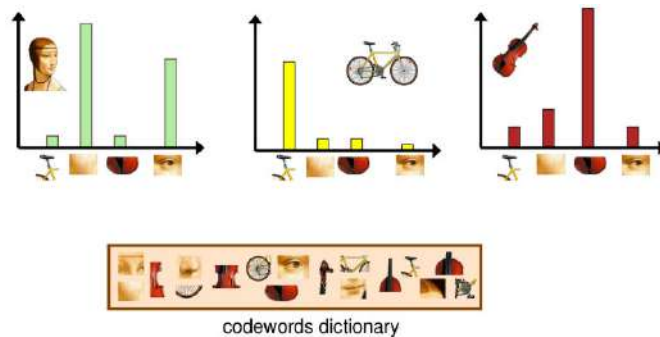


Figure 7. Codeword Dictionary[13]

E. Classification : For classification, we will be applying **Logistic regression and SVM** on visual bag of words feature vector.

Logistic Regression :

Logistic regression uses hypothesis where $g(z)$ function $g(z)$ is called the logistic function or the sigmoid function. As z tends towards 1, and as z tends $z \rightarrow \infty, g(z) \rightarrow 1, g(z) \rightarrow 0$ towards 0.

$$h\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

[15]

SVM

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In two dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side.

$$K(x, z) = \exp\left(-\frac{1}{2\tau^2} \|x - z\|_2^2\right)$$

[15]

Support vector classifier: Parameter C and are fine tunes using parameter grid search in sklearn

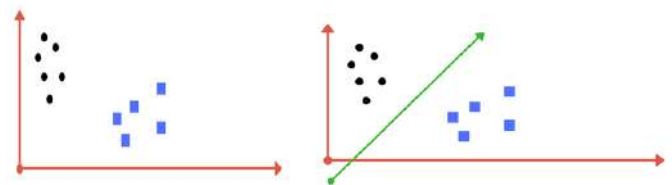


Figure 8. SVM Classification[17]

3 Requirement Analysis

The implementation detail is given in this section.

3.1 Software

The proposed approach starts by pre-processing image into gray scale images then resizing and cropping them. SIFT (Scale-invariant feature transform) computer vision algorithm will be applied on pre-processed image to detect feature descriptors in the image. Visual bag of words will be constructed from feature descriptors obtained from the images. Computed visual bag of words will be used as a feature vector for Logistic regression and SVM. Each model's output will be a binary label for prediction of each diseases.

3.3 Dataset and Parameters

Dataset has been recently released, that contains 112, 120 frontal-view X-ray images of 30,805 unique patients, with each image labeled with up to 14 lung diseases. Each image is a gray scale image with 1024 x 1024 in resolution.

Metrics used to evaluate models performance are accuracy, precision, recall, and ROC curve. Number of cluster centroids for each of the classifier is determined using accuracy and recall. With more importance to recall, because of medical domain.

ACKNOWLEDGMENT

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

REFERENCES

- 1) *Chest X-RAY Analysis to detect Mass Tissues in Lungs*, 3rd INTERNATIONAL CONFERENCE X-Ray ON INFORMATICS, ELECTRONICS & VISION 2014, Emon Kumar Dey, Hossain Muhammad Muctadir .
- 2) *Image Segmentation for Lung Region in Chest X-ray Images using Edge Detection and Morphology*, 2014 IEEE International Conference on Control System, Computing and Engineering, 28 - 30 November 2014, Penang, Malaysia, Zurina Muda, Noraidah Sahari, Hamzaini Abdul Hamid.
- 3) *Foreign Object Detection in Chest X-rays*, 2015 IEEE International Conference on Bioinformatics and Biomedicine (BTBM), Zhiyun Xue, Serna Candemir, Sameer Antani, L. Rodney Long, Stefan Jaeger, Dina Demner-Fushman, George R. Thoma Lister Hill National Center for Biomedical Communications National Library of Medicine Bethesda, USA
- 4) *Thorax Disease Diagnosis Using Deep Convolutional Neural Network*, Jie Chen, Member, IEEE, Xianbiao Qi, Osmo Tervonen, Olli Silvén, Guoying Zhao, senior Member, IEEE and Matti Pietikäinen, Fellow, IEEE, 2016
- 5) *Automatic Detection of Major Lung Diseases Using Chest Radiographs and Classification by Feed - forward Artificial Neural Network*, 1st IEEE International Conference on Power Electronics, Intelligent Control and Energy (ICPEICES-2016), Shubhangi Khobragade, Aditya Tiwari, c.Y. Patil and Vikram Narke
- 6) *Detection of Pneumonia clouds in Chest X-ray using Image processing approach*, Abhishek Sharma, Daniel Raju, Sutapa Ranjan ©2017 IEEE
- 7) *ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases*, Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, Ronald M. Summers Department of Radiology and Imaging Sciences, Clinical Center, 2 National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20892, arXiv: 1705.02315v5[cs.CV] 14 DEC 2017.
- 8) *Effective Pneumothorax Detection for Chest X-Ray Images Using Local Binary Pattern and Support Vector Machine*, Yuan-Hao Chan, 1 Yong-Zhi Zeng, 2 Hsien-Chu Wu ,Ming-Chi Wu, and Hung-Min Sun , Hindawi Journal of Healthcare Engineering Volume 2018, Article ID 2908517.
- 9) *Histogram equalization* - wikipedia. https://en.wikipedia.org/wiki/Histogram_equalization
- 10) *GridsearchCV* sklearn. http://scikitlearn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
- 11) *sklearn machine learning algorithm library*. <http://scikit-learn.org/> [5] *Visual bag of words*. <https://kushalvyas.github.io/BOV.html>
- 12) *Opencv*. https://docs.opencv.org/3.1.0/da/df5/tutorial_py_sift_intro.html
- 13) <https://kushalvyas.github.io/BOV.html>
- 14) <http://scikit-learn.org/stable/modules/svm.html>
- 15) <http://cs229.stanford.edu/proj2017/final-posters/5136599.pdf>.
- 16) <https://www.cs.ubc.ca/~lowe/papers/ijcv04.pdf>.
- 17) <https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812ef7c72>

Prediction of Indian Election Sentiments on Twitter using Machine Learning

Rohith Nair, Sharan Rai, Vickson Rodrigues, Shivam Soni, Manasi Kulkarni
Department of Computer Engineering, PCE, Navi Mumbai, India - 410206

Abstract— *Sentiment analysis is considered to be a category of machine learning and natural language processing. It is used to extract and recognize opinions from different content structures, including news, audits and articles and categorizes them as positive, neutral and negative. The main objective here is to provide insights about the public opinion about different political parties and predict the polarity of opinions towards different political parties. In this proposed implementation we perform sentiment analysis of opinions and views on political parties and candidates posted on Twitter, a microblogging service. For this project we aim to explore deep learning techniques such as RNN (Recurrent Neural Network) and CNN (Convolutional Neural Network) and do a comparative study with traditional machine learning algorithms used for sentiment analysis such as SVM.*

Keywords—Sentiment analysis, Natural Language processing, Machine Learning and RNN (Recurrent Neural Network) SVM (Support Vector Machine)

I. INTRODUCTION

Sentiment analysis is considered to be a category of machine learning and natural language processing. It is used to extricate, recognize, or portray opinions from different content structures, including news, audits and articles and categorizes them as positive, neutral and negative. Our aim is to apply sentiment analysis on tweets gathered from Twitter. As Twitter is a popular micro-blogging social media platform, many people express their likes or dislikes for a political party. We use RNN (Recurrent Neural Network) which is traditional deep learning technique to calculate the sentiment

of political tweets in the data corpus collected from twitter. The result of the analysis is displayed to the user using a graphical representation with the help of android application.

II. PROPOSED WORK

The proposed system performs sentiment analysis on the data collected to predict the general sentiment of the public towards each political parties over the period of time leading upto the elections. The system aims to perform aspect based analysis to understand the subjects the tweets are about, the subject could be any of the policies by the political parties which would help to analyse how the policies or ideas that the political parties propose to implement are received by the public. The system shows how the opinion of the voters change with each major event during the respective campaigns of the political parties, the events could be rallies organised by the political parties or debates participated by the party candidates. The tweets will be represented using multidimensional vectors generated using Word2Vec model and a comparative study on the different algorithms such as RNN CNN and SVM would be done to determine the best algorithm for sentiment analysis on tweets.

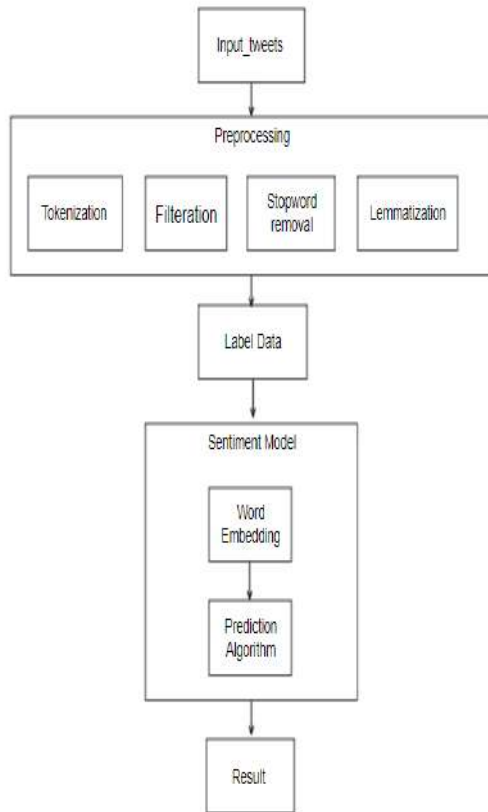


Fig: Prediction of Indian Elections

a) Input

The input here are the tweets related to the 2014 Indian elections collected from twitter.

b) Preprocessing

The data retrieved from twitter is in json form it needs to be converted into tabular form and the data needs to be cleaned to input for further processing. Some of the steps involved in preprocessing are detailed below.

c) Tokenization

Tokenization is the act of breaking up a sequence of strings into pieces such as words, keywords, phrases, symbols and other elements called tokens. Tokens can be individual words, phrases or even whole sentences. In the process of tokenization, some characters like punctuation marks are discarded.

Algorithm:-

Input: A single sentence

Output: A list of tokens

Input

No CM in Indias history has tried harder to bring his Govt down Success at last ArvindKejriwal can now get down to Lok Sabha campaign

Output

“No” “CM” “in” “India’s” “history” “has” “tried” “harder” “to” “bring” “his” “Govt” “down” “Success” “at” “last” “ArvindKejriwal” “can” “now” “get” “down” “to” “Lok” “Sabha” “campaign”

Filteration

This is done to remove the special characters(@,!,&,\$ etc) and numbers as they don’t convey much information.

Algorithm:-

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

Input:-

RT @Sumit_Nagpal: The real worry of BJP & Congress is not what if @ArvindKejriwal becomes the CM, their worry is what if he delivers what h...

Output:-

RT Sumit_Nagpal The real worry of BJP Congress is not what if ArvindKejriwal becomes the CM their worry is what if he delivers what h

e) StopWords Removal

Stop words are words which are filtered out before or after processing of natural language data. Though "stop words" usually refers to the most common words in a language, there is no single universal list of stop words used

by all natural language processing tools, and indeed not all tools even use such a list. Some tools specifically avoid removing these stop words to support phrase search.

Sample Stopword List

{ 'he', 'between', 'yourself', 'but', 'again', 'there', 'about', 'once', 'and', 'on', 'very', 'having', 'will', 'they', 'own', 'an', 'be', 'some', 'for', 'do', 'any', 'yours', 'such', 'into', 'of', 'most', 'itself', 'other', 'off', 'is', 's', 'am', 'or', 'who', 'as', 'from', 'him', 'each', 'the', 'themselves', 'until', 'below' }

Algorithm:-

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

Input

suchetadalal by far the best analysis on gas pricing and exposing kejriwal shoot and scoot wonder he will answer any

Output

suchetadalal by far best analysis gas pricing exposing kejriwal shoot scoot wonder answer

d) Lemmatization

Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the *lemma*.

English lemma list:

have -> had,has,'ve,having,'s,'d,of,d,ve
it -> its,they
he-> his,him,they
i -> my,me,we,is
they -> their,them,'em
you -> your,ya,ye
not -> n't
she -> her
do -> did,does,done,doing,du,d'

Algorithm

1. Input
2. if words word in sentence == lemma list
then goto step-4
3. else message("already in lemma form")
then goto step-4
4. output
5. Exit

Input

BDUTT AamAadmiParty Corruption the main issue AAP is fighting for has been rampant in congress regime NOT BJP AK losing my trust respect

Output

BDUTT AamAadmiParti Corrupt the main issu AAP is fight for ha been rampant in congress regim NOT BJP AK lose my trust respect

f) Word Embeddings

A word embedding is an approach to provide a dense vector representation of words that capture something about their meaning. Word embeddings are an improvement over simpler bag-of-word model word encoding schemes like word counts and frequencies that result in large and sparse vectors (mostly 0 values) that describe documents but not the meaning of the words. Word embeddings work by using an algorithm to train a set of fixed-length dense and continuous-valued vectors based on a large corpus of text. Each word is represented by a point in the embedding space and these points are learned and moved around based on the words that surround the target word. It is defining a word by the company that it keeps that allows the word embedding to learn something about the meaning of words. The vector space representation of the words provides a projection where words with similar meanings are locally clustered within the space. The use of word embeddings over other text representations is one of the key

methods that has led to breakthrough performance with deep neural networks on problems like machine translation. Here we will use word2vec embedding method to convert the textual data into multidimensional vectors which is created by google using millions of wikipedia documents.

g) Prediction Algorithm

1. RNN

A recurrent neural network (RNN) is a class of artificial neural network where connections between nodes form a directed graph along a sequence. This allows it to exhibit temporal dynamic behavior for a time sequence. Unlike feed forward neural networks, RNNs can use their internal state (memory) to process sequences of inputs.

2. LSTM

Long short-term memory (LSTM) networks were discovered by Hochreiter and Schmidhuber in 1997 and set accuracy records in multiple applications domains. Around 2007, LSTM started to revolutionize speech recognition, outperforming traditional models in certain speech applications. In 2009, a Connectionist Temporal Classification (CTC)-trained LSTM network was the first RNN to win pattern recognition contests when it won several competitions in connected handwriting recognition. In 2014, the Chinese search giant Baidu used CTC-trained RNNs to break the Switchboard Hub5'00 speech recognition benchmark without using any traditional speech processing methods.

3. SVM

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal

hyperplane which categorizes new examples. In two dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side.

4. CNN

In machine learning, a network (CNN, or ConvNet) is a class of deep, feed-forward artificial neural networks, most commonly applied to analyzing visual imagery. CNNs use a variation of multilayer perceptrons designed to require minimal preprocessing. They are also known as shift invariant or space invariant artificial neural networks (SIANN), based on their shared-weights architecture and translation invariance characteristics. Convolutional networks were inspired by biological processes in that the connectivity pattern between neurons resembles the organization of the animal visual cortex. Individual cortical neurons respond to stimuli only in a restricted region of the visual field known as the receptive field. The receptive fields of different neurons partially overlap such that they cover the entire visual field. CNNs use relatively little pre-processing compared to other image classification algorithms. This means that the network learns the filters that in traditional algorithms were hand-engineered. This independence from prior knowledge and human effort in feature design is a major advantage.

h) Result

The result is shown for the user in the form of an android app that lets the user know about the different trends about the election and the prediction about the election.

III. IMPLEMENTATION DETAILS

The tweets are collected using a pre-existing collection of tweet id's. The tweet id's are passed to a tweet hydrator app which fetches the tweet corresponding to each tweet id. The hydrated tweets are returned in json form

containing the tweet text and many other details. The required details are extracted from the json tweet output. The data is then segregated according to the respective political party and the required cleaning is done. The dataset is then labelled. A time series split of data is done into training and validation data. The training data is fed into the classifier model in the form of multidimensional vectors created using word2vec model. The model then is used for the prediction.

a) Sample Dataset

Created_at	Tweet_id	Full_text
Sat Mar 15 18:10:12 +0000 2014		@SushmaSwarajbj p sushma ji jehan per samman naa mile , vehan brahmano ko nahi rehna chahiye
Tue May 27 02:50:30 +0000 2014	61126132	RT @SriSri: Blessings & Best Wishes to Narendra Modi & his team of Ministers.May God give them the strength & wisdom to fulfil the high hop...
Sat May 10 16:13:48 +0000 2014	18839785	RT @narendramodi: I am overwhelmed by people's response! I assure them we will repay their affection with unprecedented development.

Wed Jul 02 03:02:05 +0000 2014	40542703 5	@ArvindKejriwal right action wud b 2 demand action again thieves of previous regime n keep note of present ones for next regime.
Sun May 18 20:06:24 +0000 2014	24705126	@ShashiTharoor if u wer so knowledgeable abt foreign policies, today ur govt wudnt hv faced such defeat @BDUTT

IV. REQUIREMENT ANALYSIS

a) Software Requirements:

Python - python is an interpreted, object oriented high level programming language. It emphasizes code readability and reduces code maintenance making it suitable for Rapid Application Development.

TextBlob, Spacy - Python libraries for dealing with textual data.

Scikit learn - Python library for analysis of data

Keras - Python wrapper for using Tensorflow.

Android,flask - For creating android dashboard.

b) Hardware Requirements

Google Cloud Platform - For working with large amounts of data

V. CONCLUSION

A comparative study on different algorithms used for sentiment analysis on tweets is done. Most of the current implementations have

seldom explored deep learning algorithms like RNN (Recurrent Neural Networks) and CNN (Convolutional Neural Networks) here we look to see how neural networks compare with the traditional algorithms like SVM (Support Vector Machine). The dataset contains about 22 million tweets which are used for analysis and prediction of the sentiment of the general twitter users towards each political party and how this reflects the sentiment of the general population. Volume analysis is performed to recognize the different trends found with the frequency of the tweets. The results are shown in the form of an interactive android application.

VI. ACKNOWLEDGMENT

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

VII. REFERENCES

[1] B. Joyce and J. Deng, "Sentiment analysis of tweets for the 2016 US presidential election," *2017 IEEE MIT Undergraduate Research Technology Conference (URTC)*, Cambridge, MA, 2017, pp. 1-4.

[2] P. Sharma and T. Moh, "Prediction of Indian election using sentiment analysis on Hindi Twitter," *2016 IEEE International Conference on Big Data (Big Data)*, Washington, DC, 2016, pp. 1966-1971.

[3] Yoon Kim. Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882, 2014.

[4] Abhishek Bhola "Twitter and Polls: Analyzing and estimating political orientation of Twitter users in India General Elections 2014" arXiv:1406.5059 [cs.SI]

[5] A. Timmaraju, V. Khanna, "Sentiment analysis on movie reviews using recursive and recurrent neural network architectures", *Semantic Scholar*, 2015.

[6] Socher, R & Perelygin, A & Wu, J.Y. & Chuang, J & Manning, C.D. & Ng, A.Y. & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. EMNLP. 1631. 1631-1642.

[7] H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan, "A system for real-time twitter sentiment analysis of 2012 u.s. presidential election cycle," in Proceedings of the ACL 2012 System Demonstrations, ACL '12, (Stroudsburg, PA, USA), pp. 115–120, Association for Computational Linguistics, 2012.

[8] Tumasjan A, Sprenger TO, Sandner PG, Welpe IM. Election forecasts with Twitter: How 140 characters reflect the political landscape. Soc Sci Comput Rev. 2011;29:402–418.

[9] Pak, A. & Paroubek, P., 2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. Valetta, s.n., pp. 1320-1326

Assist Crime Prevention Using Machine Learning

Nair Swati Sasindrakumar, *Student, PCE*, Soniminde Saloni Ajit, *Student, PCE*, Sruthi Sureshababu, *Student, PCE*, Apurva Chandrakant Tamhankar, *Student, PCE* and Sagar Kulkarni, *Faculty, PCE*

Abstract—Crime rate is increasing significantly over the years. Crime prevention is an attempt to reduce and deter the crime rates and number of criminals. The government must go beyond law enforcement and criminal justice to tackle the risk factors that cause crime because it is more cost effective and leads to greater social benefits than the standard ways of responding to crimes. It has been observed that criminals follow a certain pattern. The data driven method is used which is based on the broken windows theory, having an enormous impact on the working of the police department. The theory links disorder and incivility within a community to subsequent occurrences of serious crimes. Predictive model for crime is developed using Machine learning. Predictive policing is used by the law enforcement stakeholders to decide on how to allocate resources, manage and successfully avoid situations by taking proactive measures against thefts, robberies, homicides and other crimes. This will help the bureau and the police departments to efficiently focus their resources on locations which are potential crime hotspots. Data driven approach is needed to automate the prediction process by identifying the interconnections and pattern in the data. The model is built using this approach to predict the crime rate based on demographic and economic information of particular localities using decision trees, linear classification, regression and spatial analysis.

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

I. INTRODUCTION

Crimes are increasing day by day which means that there should be measures to avoid them. Crime prevention refers to recognizing that a crime risk exists and taking some corrective action to eliminate or reduce that risk. Using machine learning approach we will assist the local authorities in preventing crime and to take the necessary actions against crime.

There are numerous types of crimes taking place at different locations. Some areas have crimes occurring frequently whereas there are some places where occurrence of crime is negligible. Therefore potential crime hotspot areas require much more security than those areas where crime rate is comparatively less. For example, Crimes like chain snatching occur mostly at lonely places so that criminals could escape easily from that location. Detecting the crime hotspot areas helps the police officials to decide what kind of security

strength will be required for that particular place.

The system is based on the broken windows theory. Broken windows theory is an academic theory proposed by James Q. Wilson and George Kelling in 1982 that used broken windows as a metaphor for disorder within neighbourhoods. Their theory links disorder and incivility within a community to subsequent occurrences of serious crime [12].

II. INFERENCE

Crime Prevention has been done using data mining and machine learning techniques but in this system combination of machine learning algorithm is used comparing the accuracy obtained by each method, best result is provided. This system also includes geo-spatial analysis and user interface to detect various types of crime occurrences at a given place whereas other systems are used for detecting only a specific kind of crime or crimes following a specific pattern that is modus operandis.

III. SCOPE OF THE PROJECT

The scope of the crime prevention system is to assist the police department and bureau in predicting the crime before it's occurrence by conducting analysis on the previous data using Machine Learning. Machine Learning models will help in classifying with minimal error whether an area is a potential crime hotspot or not based on the community crime dataset. The broken window theory is adopted. The accuracy and precision of the models is calculated to evaluate the performance of the system. The output will indicate whether the area is a potential crime hotspot or not. If it is then the police department will have to deploy more security forces in that area to prevent crime from happening.

IV. RELATED WORK

We have cited the relevant past literature of research work done in the field of crime prevention for different locations.

[1] **Ying-Lung Lin, Liang-Chih Yu, Tenge-Yang Chen** presented a method where Predictions are based on a period of one month. The Map is split into various grids combined with 56 features. Method used is spatial analysis. This model was specific to Taiwan crime record.

[2] **Devendra Kumar Tayal, Arti Jain, Surbhi Arora, Surbhi Agarwal, Tushar Gupta, Nikhil Tyagi** presented an approach for the design and implementation of

crime detection and criminal identification for Indian cities using data mining techniques. It is divided into six modules, namely—data extraction (DE), data preprocessing (DP), clustering, Google map representation, classification and WEKA implementation.

V. SYSTEM ARCHITECTURE

System architecture shows the overall flow of the System.

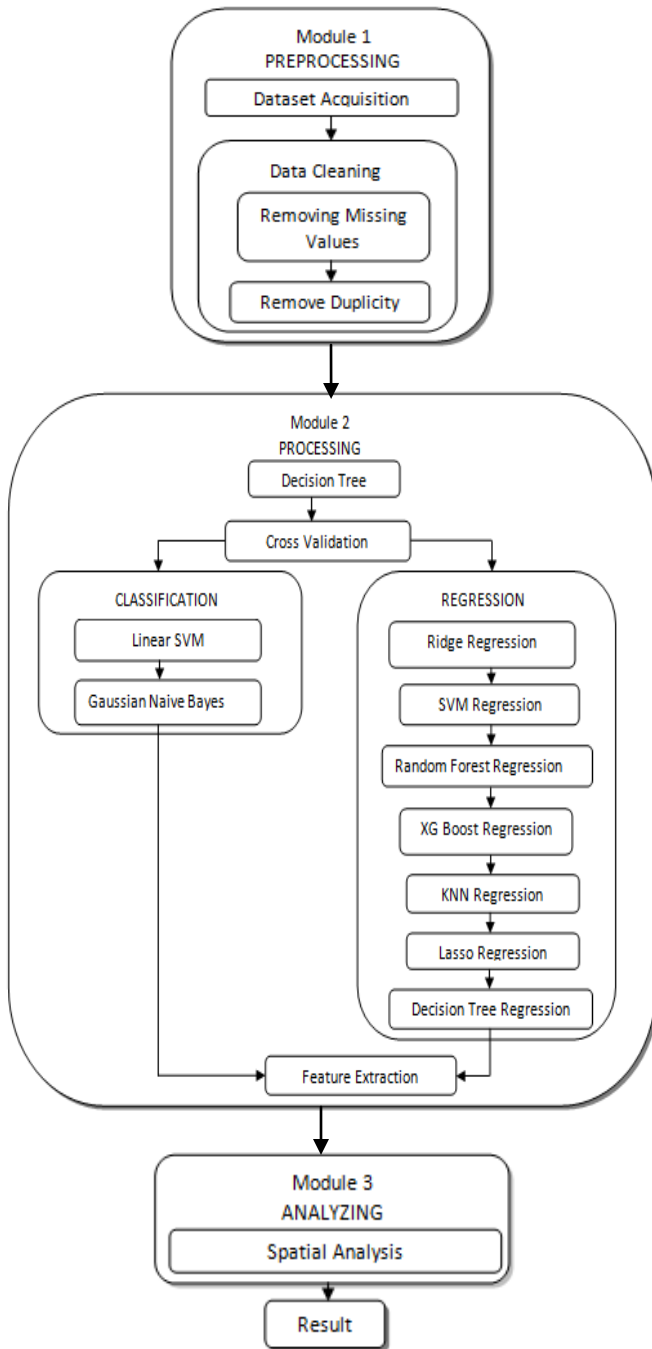


Figure1: System architecture

VI. METHODOLOGY

System works on following stages.

A. Preprocessing

Pre-processing is the process of cleaning and preparing the text for classification. In this phase the input text is completely filtered so that unwanted characters and symbols are removed from the text. The missing values in data set are cleaned and the data is made appropriate for further processing

Algorithm for Pre-processing module:

1. Accept the data set in csv format (comma separated value file)
2. Remove corrupt data.
3. Impute missing data.

The communities-crime-full.csv dataset is used. The dataset consists of the crime records of the communities within the United States.

Table 1: Dataset before cleaning:

	A	B	C	D	E	F	G	H	
1	state	county	communit	communit	fold	populatio	householi	racepctbl	race
2		8 ?	?	Lakewood		1	0.19	0.33	0.02
3		53 ?	?	Tukwila	cit	1	0	0.16	0.12
4		24 ?	?	Aberdeen		1	0	0.42	0.49
5		34	5	81440	Willingbo	1	0.04	0.77	1
6		42	95	6096	Bethleher	1	0.01	0.55	0.02
7		6 ?	?	SouthPas		1	0.02	0.28	0.06
8		44	7	41500	Lincolntov	1	0.01	0.39	0
9		6 ?	?	Selmacity		1	0.01	0.74	0.03
10		21 ?	?	Henderso		1	0.03	0.34	0.2
11		29 ?	?	Claytoncit		1	0.01	0.4	0.06
12		6 ?	?	DalyCityci		1	0.13	0.71	0.15
13		36 ?	?	Rockvillec		1	0.02	0.46	0.08
14		25	21	44105	Needham	1	0.03	0.47	0.01
15		55	87	30073	GrandChu	1	0.01	0.44	0

The dataset is for per-capita crime rates around the country. Our task is to build models to predict the crime rate based on demographic and economic information about the particular locality. The data is given in the file “communities-crime-full.csv”. It includes data fields with missing values (indicated by “?”), which have to be removed.

Table 2: Dataset after cleaning

	A	B	C	D	E	F	G	H	
1	state	communit	fold	populatio	householi	racepctbl	racePctW	racePctAs	raceI
2		1 Alabast	7	0.01	0.61	0.21	0.83	0.02	
3		1 Alexander	10	0.01	0.41	0.55	0.57	0.01	
4		1 Annistonc	3	0.03	0.34	0.86	0.3	0.04	
5		1 Athenscit	8	0.01	0.38	0.35	0.71	0.04	
6		1 Auburncit	1	0.04	0.37	0.32	0.7	0.21	
7		1 Bessemer	6	0.04	0.44	1	0.1	0	
8		1 Birminghs	2	0.41	0.37	1	0.02	0.03	
9		1 Cullmanci	1	0.01	0.3	0	0.99	0.02	
10		1 Daphnecl	7	0	0.39	0.31	0.75	0.02	
11		1 Decaturcit	10	0.06	0.39	0.32	0.73	0.04	
12		1 Dothanct	4	0.07	0.41	0.53	0.57	0.05	
13		1 Enterpris	2	0.02	0.43	0.41	0.64	0.08	
14		1 Eufulacit	5	0.01	0.46	0.67	0.47	0.02	
15		1 Fairfieldci	1	0	0.45	1	0	0.01	

B. Processing:

1. Decision tree

The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. The decision rules are generally in form of if-then-else statements. The deeper the tree, the more complex

the rules and fitter the model. We will use the clean dataset to predict whether the crime rate in a locality is greater than 0.1 per capita or not. A new field “highCrime” is created which is true if the crime rate per capita (ViolentCrimesPerPop) is greater than 0.1, and false otherwise. The percentages of positive and negative instances in the dataset are found. The DecisionTreeClassifier is used to make a decision tree learn to predict highCrime on the entire dataset. The training accuracy, precision, and recall for this tree is then calculated. The main features used for classification are later identified.

2. Cross Validation

Cross-validation is a statistical method used to estimate the skill of machine learning models. Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation.

Algorithm for Cross Validation:

1. Shuffle the dataset randomly.
2. Split the dataset into k groups
3. For each unique group:
 - a. Take the group as a hold out or test data set
 - b. Take the remaining groups as a training data set
 - c. Fit a model on the training set and evaluate it on the test set
 - d. Retain the evaluation score and discard the model
 - e. Summarize the skill of the model using the sample of model evaluation scores
 - f. Importantly, each observation in the data sample is assigned to an individual group and stays in that group for the duration of the procedure. This means that each sample is given the opportunity to be used in the hold out set 1 time and used to train the model k-1 times.

We will apply cross-validation (cross_val_score) to do 10-fold cross-validation to estimate the out-of-training accuracy of decision tree learning. We will find out what are the 10-fold cross-validation accuracy, precision, and recall.

3. Classification

In machine learning, classification is the problem of identifying to which set of categories a new observation belongs, on the basis of a training set of data containing observations whose category membership is known.

- Linear SVM

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. The LinearSVC is used to make a linear Support Vector Machine model learn to predict highCrime.

- i. The 10-fold cross-validation accuracy, precision, and recall

for this method is found.

- ii. The 10 most predictive features are identified.
- iii. The results are the compared with results from decision trees.

- Gaussian Naive Bayes

Bayes’ Theorem provides a way that we can calculate the probability of a hypothesis given our prior knowledge. Bayes’ Theorem is stated as:

$$P(h|d) = (P(d|h) * P(h)) / P(d) \quad \dots(1)$$

Where,

P(h|d) is the probability of hypothesis h given the data d. This is called the posterior probability.

P(d|h) is the probability of data d given that the hypothesis h was true.

P(h) is the probability of hypothesis h being true (regardless of the data). This is called the prior probability of h.

P(d) is the probability of the data (regardless of the hypothesis)

The GaussianNB is used to make a Naive Bayes classifier learn to predict highCrime.

- i. The 10-fold cross-validation accuracy, precision, and recall for this method is found.
- ii. The 10 most predictive features are identified.
- iii. The results are the compared with results from decision trees.

4. Regression

Regression is used to predict continuous values. We perform regression analysis to understand which among the independent variables are related to the dependent variable. [11]

Examples of regression are:-

- Verifying the relationship between house pricing vs a whole bunch of exogenous variables such as neighbourhood, location, bathrooms in the house, bedrooms, how far is it from the city.
- Estimate the relationship between the stock market index and it's relationship with the macro economic variables.
- Examining the exchange rate movement and its dependency on several key macro economic factors.

Regression will be used for predicting the crime rate per capita (ViolentCrimesPerPop). The following errors are calculated:

1. RMSE(Root Mean Square Error)
2. MAE(Mean Absolute Error)
3. R²(R Square Error)

- Ridge Regression

Ridge Regression is a technique for analyzing multiple regression data that suffer from multicollinearity. When

multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors.

- SVM Regression

SVM constructs a hyperplane in multidimensional space to separate different classes. SVM generates optimal hyperplane in an iterative manner, which is used to minimize an error.

- Random Forest Regression

It is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

- XGBoost Regression

XGBoost stands for eXtreme Gradient Boosting. It is an implementation of gradient boosted decision trees designed for speed and performance.

- KNN Regression

K nearest neighbors is a simple algorithm that stores all available cases and predict the numerical target based on a similarity measure

- Lasso Regression

Lasso (Least Absolute Shrinkage and Selection Operator) penalizes the absolute size of the regression coefficients. In addition, it is capable of reducing the variability and improving the accuracy of linear regression models.

- Decision Tree Regression

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

The best results are obtained from Random Forest and XGBoost Regression depending upon the input given to the model.

Algorithm for predicting crime

1. Taking input dataset which is .csv file (In our example we have US based dataset).
2. Perform cleaning and pre-processing. Save the cleaned file and use this for further analysis.
3. Based on various conditions, apply appropriate decision tree and infer the results.
4. Split the data into train and test by using cross validation.
5. Apply various Classification and Regression models. Analyze them using evaluation metrics and select one which gives best results.

6. Perform spatial analysis using GeoPanda library.
7. Based on the results obtained we can identify the area of high crime and inform authority in order to prevent crime from happening.

5. Feature Extraction

Determining a subset of the initial features is called feature selection. The selected features are expected to contain the relevant information from the input data, so that the desired task can be performed by using this reduced representation instead of the complete initial data. It involves reducing the amount of resources required to describe a large set of data. When performing analysis of complex data one of the major problems stems from the number of variables involved.

C.Spatial Analysis:

Spatial analysis is a type of geographical analysis which seeks to explain patterns of human behavior and its spatial expression in terms of mathematics and geometry, that is, locational analysis.

GeoPandas is the geospatial implementation of the big data oriented Python package called Pandas. GeoPandas enables the use of the Pandas data types for spatial operations on geometric types. The potential crime hotspots are plotted on the map which gives better visualization of results.

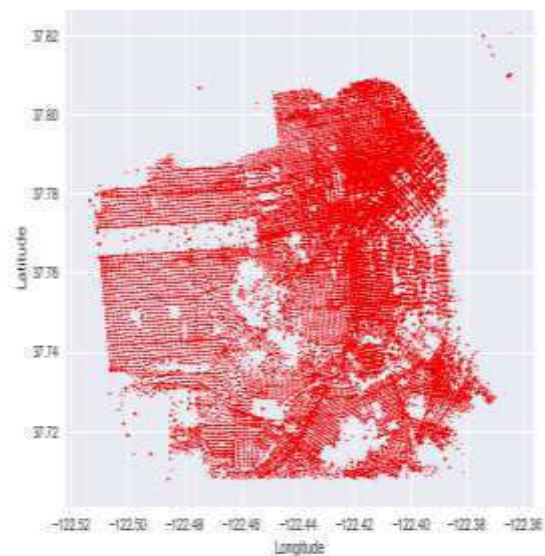


Figure 2: Plot of Crime Hotspots

VII. RESULT ANALYSIS

The quality of a domain system can be evaluated by comparing recommendations to a test set of known user ratings. These systems are typically measured using accuracy, precision and recall.

Table 3: Prediction Outcomes

Condition Positive (P)	The number of real positive cases in the data
Condition Negative (N)	The number of real negative cases in the data
True Positive (TP)	Equivalent to hit
True Negative (TN)	Equivalent to correct rejection
False Positive (FP)	Equivalent to false alarm, Type I error
False Negative (FN)	Equivalent to miss, Type II error

Precision: A measure of exactness, determines the fraction of relevant items retrieved out of all items retrieved. Precision (P) It is given in Equation 2.

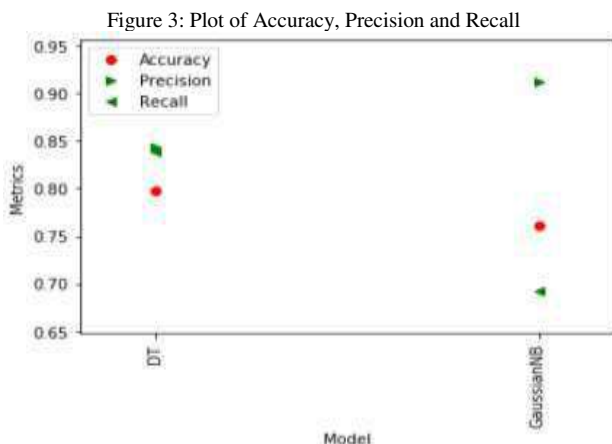
$$P = \frac{TP}{TP + FP} \dots(2)$$

Accuracy: Accuracy is the proximity of measurement results to the true value; precision, the repeatability, or reproducibility of the measurement. Accuracy (A) is given in Equation 3.

$$A = \frac{TP + TN}{P + N} \dots(3)$$

Recall: a measure of completeness, determines the fraction of relevant items retrieved out of all relevant items. Recall (R) is given in Equation 4.

$$R = \frac{TP}{TP + FN} \dots(4)$$



R squared error (R²): It is a statistical measure of how close the data are to the fitted regression line. R² is given in Equation 5.

$$R^2 = 1 - \frac{\text{Sum Squared Regression Error}}{\text{Sum Squared Total Error}} \dots(5)$$

Root Mean Square Error (RMSE): It is a frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed. RMSE is given in Equation 6.

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (\text{Predicted}_i - \text{Actual}_i)^2}{N}} \dots(6)$$

Mean Absolute Error (MAE): It is a measure of difference between two continuous variables. Consider a scatter plot of *n* points, where point *i* has coordinates (*x_i*, *y_i*). Mean Absolute Error is the average vertical distance between each point and the identity line. MAE is given in Equation 7.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - x_i| \dots(7)$$

VIII. APPLICATIONS

Technical applications

- Assist police department for crime prevention
The Crime Prevention System will assist police department in maintaining law and order, as the model will give a pictographic view of crime hotspots based on the data set provided of that region.

- Crime Reports for newspapers
This system can be used by news reporters or journalists to give a brief analysis about crime occurrences at a particular place stating about the type of crime and its frequency.

- Predicting crimes from news feeds
Crime patterns can be analyzed and crimes can be predicted from news feeds. The news feeds for a particular time span can be collected like for 20 years and this news feeds corpus can be used to predict future events.

Social Applications

- Combat drug addiction and other related crime
This system will help to identify the predominant drug and other related crime hotspots and then the government can set up rehabilitation centres and camps. Non-governmental organizations (NGOs) can also conduct awareness programmes for the same.

- Urban planning

Once the crime hotspots are identified the government can take measures to redevelop those areas by implementing urban planning so as to improve the social neighbourhood of a person by which there is no or minimal indulgence in criminal or illegal activities. Bad urban planning can lead to an increase in crime rate.

- Analyzing crime through social media

The tweets and social media posts can be analyzed for a certain timespan. From this corpus certain deductions can be made about the crime patterns and criminal instincts. By further enhancements on the model using Natural Language Processing, the crimes can be prevented from happening by assessment of social media posts.

IX. CONCLUSION

The project uses different Machine Learning approaches to assist in crime prevention by predicting whether a particular area is a potential crime hotspot or not. The community crimes dataset of the US is used for this purpose. As the dataset collected consists of missing values, it has to be cleaned and preprocessed. Decision trees can then be used to make decision about a high crime area. The classification models are applied to the system and the topmost features can be predicted. Different regression models are applied aiming for the least error. The model with the least error will be the winning model. Accuracy, Precision and Recall are considered for evaluation of the system. Geospatial analysis can then be done to plot the potential crime hotspots across the longitudinal and latitudinal positions over a map. This plot will assist the police department in deciding which area requires greater attention and hence larger security forces could be deployed at that particular crime hotspot.

ACKNOWLEDGEMENT

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

We are thankful to **Dr. Madhumita Chatterjee**, H.O.D, Computer Engineering Department and **Prof. Gaurav Sharma**, B.E. Project Coordinator, Pillai college of Engineering, New Panvel, for his guidance, encouragement and support during our project.

It is a great pleasure and moment of immense satisfaction for us to express our profound gratitude to our Project Guide, **Prof. Sagar Kulkarni** whose constant encouragement enabled us to work enthusiastically. Without his encouragement this paper wouldn't have been published.

REFERENCES

[1] Ayisheshim Almaw, Kalyani Kadam, "Survey Paper on Crime Prediction using Ensemble Approach ,International Journal of Pure and Applied Mathematics", Vol. 118 No. 8, Pune, India. (2018) ISSN: 1311-8080 (printed version); ISSN:

1314-3395 (on-line version)

[2] Ying-Lung Lin, Liang-Chih Yu, Tenge-Yang Chen, "Using Machine Learning to Assist Crime Prevention",Taiwan. (2017) INSPEC Accession Number: 17375465

[3] N.D. Waduge, Dr. L. Ranathunga , "Machine Learning Approaches To Detect Crime Patterns", Sri Lanka. (2017)

[4] Hyeon-Woo Kang, Hang-Bong Kang , "Prediction of crime occurrence from multimodal data using deep learning", Plos One, Bucheon, Gyonggi-Do, Korea. (2017)

[5] Lawrence McClendon and Natarajan Meghanathan, "Using Machine Learning Algorithms To Analyze Crime Data", Machine Learning and Applications: An International Journal (MLAIJ), USA. (2015)

[6] Harsha Perera, Shanika Udeshini, Malith Munasinghe, "Criminal short listing and crime forecasting based on modus operandi" , 14th International Conference on Advances in ICT for Emerging Regions (ICTer) ,Colombo,SriLanka (2014) INSPEC Accession Number: 15058519

[7] Devendra Kumar Tayal, Arti Jain, Surbhi Arora et.al., "Crime detection and criminal identification using data mining", Springer-Verlag, London. (2014) ,ISSN: 0951-5666

[8] Shiju Sathyadevan, Devan M.S, Surya Gangadharan S , "Crime Analysis and Prediction Using Data Mining" , 2014 First International Conference on Networks & Soft Computing , Guntur, India (2014) DOI: 10.1109/CNSC.2014.6906719

[9] Andrey Bogomolov , Bruno Lepri, Jacopo Staiano et.al.,"Once Upon a Crime: Towards Crime Prediction from Demographics and Mobile Data", Proceedings of the 16th International Conference on Multimodal Interaction,Istanbul, Turkey. (2014) Pages 427-434

[10] Lenin Mookiah, William Eberle and Ambareen Sira ,"Survey of Crime Analysis and Prediction" ,Proceedings of the Twenty-Eighth International Florida Artificial Intelligence Research Society Conference ,Cookeville, Tennessee(2014)

[11] <http://scikit-learn.org> , last accessed on 28th October, 2018.

[12] <https://www.britannica.com> , last accessed on 29th October, 2018.

Cyberbullying Detection & Prediction in Twitter

Najih Shafique, Sangita Tandel, Aditi Gurav, Anjana Nair, and

Prof. Madhumita Chatterjee

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

Abstract—Research into detection of cyber bullying has been increased in recent years but prediction is still on papers. Cyberbullying is found almost in all online social networks. Direct bully statements can be traced but Sarcasm is a nuanced form of communication where the individual states opposite of what is implied. One of the major challenges of sarcasm detection is the ambiguous nature. Unlike the other approaches focus on detecting the cyberbullying act based just on negative, aggressive words but also on sarcasm or those words that might lead to bullying. We will focus on those expression of sarcasm-”positive sentiments attached with negative situation”. We will use machine learning, data mining techniques to identify the characteristics of cyberbullying exchange and automatic detection of the identified traits. Approaches to be used will be Tf-idf, svm classifier, Linear Regression, Naive Bayes. Every day hundreds of new slang words are being created and used on these sites. Hence, the existing corpus of positive and negative sentiments may not prove to be accurate in detecting sarcasm. We will evaluate our methodology using tweets from different users, and show that machine learning algorithm can give accuracy in classifying the bully tweets.

Keywords— Bully, Sarcasm, Detection.

1. Introduction

Cyberbullying is defined as an aggressive, intentional action against a defenseless person by using the Internet, or other electronic contents. Researchers have found that many of the bullying cases have tragically ended in suicides; hence prediction and detection of cyberbullying has become important. In this study we show the effects of feature extraction, feature selection, and classification methods that are used, on the performance of detection of cyber bullying. To perform the experiments twitter dataset is used and the effects of preprocessing methods; several classifiers like linear regression and SVM. With the increased use of the Internet, and the ease of access to online communities provide an avenue for cyber crimes like cyber bullying. Researchers should study cyber bullying with respect to its detection, prevention and mitigation. Day by day, the effects of cyberbullying have become more serious for its victims . In many cyberbullying cases, victims have attempted suicide due to the emotionally abusive, humiliating, and aggressive messages left by predators . In the majority of cases, younger victims need to hide their predicament from adults (parents/teachers), since they think that they might lose their mobile. The challenges in fighting cyber bullying include: detecting online bullying when it occurs; reporting it to law enforcement agencies, Internet service providers and others; and identifying predators and their victims. In the literature, cyber bullying has been studied extensively from the social perspective, especially with respect to understanding its various attributes and its prevalence. However, very little attention has been focused on its online detection.

2. Literature Survey

A. Machine learning approach for detection of cyber aggression comments on social media network,(2015)[1]

In this approach[1] they have devised methods to detect cyberbullying using supervised learning techniques. They present two new hypotheses for feature extraction to detect offensive comments directed towards peers which are perceived more negatively and result in cyberbullying techniques to detect the insults and offensiveness of the comments present in social networking sites. The methodology used in this paper [1] are normalization, standard feature extraction, feature selection and finally classification. In Normalization[1] there was a removal of unwanted string and correcting words. For feature extraction two methods used are n-gram and tf-idf. Then feature selection is done on the processed data and then classifier will classify it as aggressive or not.

B. Mean Birds: Detecting Aggression & Bullying on Twitter.[2]

In their work[2] the authors have considered various machine learning algorithms, either probabilistic, tree-based, or ensemble classifiers. In this they have designed and executed a novel methodology geared to label aggressive and bullying behavior in Twitter. Their work [2] advances the state-of-art on cyberbullying and cyber aggression detection by proposing a scalable methodology for large-scale analysis and extraction of text, user, and network based features on Twitter, which has hardly been studied in this context before. They showed that their methodology for data analysis, labeling, and classification can scale up to millions of tweets, while the machine learning model built with a Random Forest classifier can distinguish between normal, aggressive, and cyberbullying users with high accuracy (> 91%). Additionally, they wanted to specifically examine the existence of hate speech and curse words within tweets. For this purpose, they have used the Hatebase database.

C. Cyberbullying Detection System on Twitter, April 2015.[3]

In their work[3], with the advent of this cyberbullying detection and solution system in Twitter, it will help the authorities to monitor, regulate or at least decrease the harassing incidents in cyberspace in Malaysia. With the implementation of the system, this will also help to raise the cyberbullying awareness among the Twitter users, and posting the tweets responsibly in the social media, as posting irritating tweets is illegal and bullies can be convicted under the Computer crimes Act, the Penal Code or the Juvenile Act, depending on the nature or severity of the case. Therefore the tweets are collected and made as a dataset and then classify each tweets as bullie or not.

D. Collaborative Detection of Cyber bullying Behavior in Twitter Data,August 2017[4]

This thesis[4] has provided a collaborative approach for detecting cyberbullying in tweets using different distributed collaboration patterns. Distributed-Collaborative approach was tested by them using experiments that were performed with three, four, and five detection nodes networks[4] that has homogeneous configurations. From the results of the three, four, and five detection nodes studies, the author concluded that the 'OR' merging technique with 2 or 3 opinions form an optimum configuration for distributed collaborative approach as it yields better recall in all the cases as compared to 'AND' and 'Majority' techniques.

E. Sarcasm Detection of Tweets: A comparative study(2017).[5]

In this paper[5] few machine learning algorithms such as Weighted Ensemble, Random Forest, logistic Regression, naive bayes are used. Here, the system also uses pragmatic classification of text in order to incorporate the role of emoticons in expressing the sentiment of text, as tweets with only positive or only negative sentiments can still be sarcastic based on emoticons associated. As nowadays emoticons plays a major role in online network, more than words people use emoticons to bully each others. So this approach basically targets on the pragmatic methodology. They also presented a novel algorithm[5] that automatically learns phrases corresponding to positive sentiments and phrases corresponding to negative situations. Using these,they learned phrases as features. With help of pragmatic classifier they were able to classify the emoticons as positive or negative.

2.1 Summary of Related Work

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

Table 1 Summary of literature survey

Literature	SVM Classifier	Tf-idf	Normalization	Pragmatic Classifier
Chavan. V.S. 2015 [1]	No	Yes	Yes	No
Nicolas Kourtellis 2017 [2]	No	Yes	No	No
Liew Chong Hong 2015 [3]	Yes	Yes	Yes	No
Amrita Mangaonkar 2017 [4]	Yes	Yes	No	No
Tanya Jain 2017[5]	No	No	Yes	Yes

3. Proposed Work

Given a set of tweets, we aim to classify each one of them depending on whether it is sarcastic or not and used to bully an user. Therefore, from each tweet, we extract a set of features, refer to a training set and use machine learning algorithms to perform the classification. The features are extracted in a way that makes use of different components of the tweet, and covers different types of sarcasm. The set of tweets on which we will run our experiments will be checked and annotated manually.

3.1 System Architecture

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

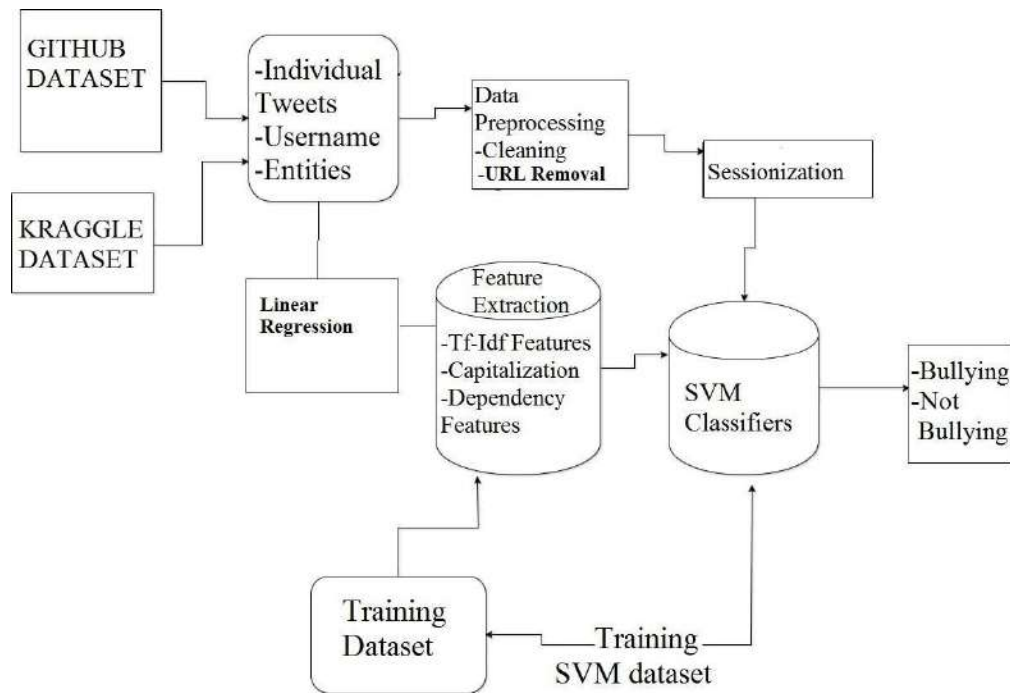


Fig. 1 Proposed system architecture

A. Input Block Description: The first part is to collect data and make two types of dataset, one will consist of plain tweets dataset and the other will consist of annotated sarcasm tweets to check whether it is bullying or not. The annotated dataset will help to predict if the user is going to be bullied or not.

B. Preprocessing: The preprocessing procedure will remove all the web links and unknown characters. Next, for each sentence, the incorrect wording is corrected. The word will be first mapped to the WordNet Lexical database. If an entry is not found, we will seek whether it has an entry in the list of saved attributes. If no entry is found at any of the attributes, we will check for the presence of character duplication that will be removed. By the help of a spell corrector, it will try converting the misspelled word into an accurate word and then the data will be processed.

C. Feature Extraction: It is the choice of the features that will be applied to the classifier. The proposed system will be using the following features:

Tf-idf (Term frequency times inverse document frequency):

TF-IDF is an information retrieval technique that weighs a term's frequency (TF) and its inverse document frequency (IDF). The TF*IDF algorithm is used to weigh a keyword in any content and assign the importance to that keyword based on the number of times it appears in the document.

The reason for using this is that bullying comments often contain bad words and scaling these features can make it easier to find a good separation for the classifier.

Capitalization:In view of social observation, words, excluding Named entities and sentence starting letters, with capitalization may convey strong relationship to cyberbullying. Therefore one counts the total number of occurrence of such capitalization in the tweets which will be helpful in predicting or detecting of cyberbullying.

Dependency Features:Whenever there will be a occurrence of bully or sarcasm words, it is related to a pronoun or to a person named-entity or username. This allow us to quantify effectively the association of sarcasm/bully words to second name/ person entity. This will help the classifier to easily trace the bully traits and notify whether bullied or not.

D. Classifier:Due to its proven efficiency in many implementation, we will be using Support Vector Machine(SVM) classifier for the system.

SVM

Support Vector Machine (SVM) Support Vector Machines (SVMs) are the newly supervised machine learning technique .SVMs revolve around the notion of a “margin”—either side of a hyperplane that separates two data classes. Maximizing the margin and thereby creating the largest possible distance between the separating hyperplane and the instances on either side of it has been proven to reduce an upper bound on the expected generalization error. The model complexity of an SVM is unaffected by the number of features encountered in the training data (the number of support vectors selected by the SVM learning algorithm is usually small). For this reason, SVMs are well suited to deal with learning tasks where the number of features is large with respect to the number of training instances. Even though the maximum margin allows the SVM to select among multiple candidate hyperplanes, for many datasets, the SVM may not be able to find any separating hyperplane at all because the data contains misclassified instances.

E. Prediction

Like detection we will also predict whether a particular tweet is going to be bullied or not. N notify before the harm is made to any individual. For prediction we use linear regression algorithm.

Linear regression:

It is a basic type of predictive analysis.

Three major uses for regression analysis are (1) determining the strength of predictors, (2) forecasting an effect, and (3) trend forecasting.

First, the regression might be used to identify the strength of the effect that the independent variable(s) have on a dependent variable.

Second, it can be used to forecast effects or impact of changes. That is, the regression analysis helps us to understand how much the dependent variable changes with a change in one or more independent variables.

Third, regression analysis predicts trends and future values. The regression analysis can be used to get point estimates.

These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula

$$“y = c + b*x”$$

where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

F.Sessionization:

It is analyzing single tweets does not provide enough context to discern if a user is behaving in an aggressive or bullying way, we group tweets from the same user, based on time clusters, into sessions and analyze them instead of single tweets.

G. Output Block Description: After cyber bully being classified or predicted, the system will notify the user abt the bully status when detected. And if predicted prior that the user is gonna be bullied then send a alert.

3 Requirement Analysis

The implementation detail is given in this section.

3.1 Software

The proposed methodology is based on the tweets posted by user so for maintaining the tweets we will require a database MySQL 5.1 & Above. For implementing the backend of the system we will use Python and PTK tool. For few Algorithm we will use java language and toolkit will be JDK 1.7 & Eclipse IDE. For the Server we will be using Apache Tomcat. In detection and prediction of cyberbully traits we will require machine learning algorithm, so will use weka toolkit to implement.

3.2 Hardware

The hardware requirement for the proposed system will be a processor above 1.9GHz to process the whole system. RAM of 2GB and 10GB Hard Disk. For input devices will be Standard keyboard and mouse. For output device it will be VGA and high Resolution Monitor.

3.3 Dataset and Parameters

Firstly we will collect dataset from two different website. One for the plain tweets dataset and the other sarcasm dataset which will be annotated. An SQL server database will be assigned in order to store and index all the database attributes, which will ultimately boost the indexing and retrieving function.

ACKNOWLEDGMENT

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

REFERENCES

- [1]“Machine learning approach for detection of cyber aggressive comments by peers on social media network”, Chavan V. S., & Shylaja, S., In Advances in computing, communications and informatics (ICACCI), 2015 International Conference on (pp. 2354-2358),2015.
- [2]“Mean Birds: Detecting Aggression and Bullying on Twitter”, Despoina Chatzakou, Nicolas Kourtellis, arXiv:1702.06877v3 [cs.CY] 12 May 2017.
- [3]“Cyberbullying Detection System on Twitter”, Liew Choong Hon, Kasturi Dewi Vara2015, Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia. vol 1(no 1) ,ISSN- 2289-2265, April 2015.
- [4]“Collaborative detection of Cyber Bullying behaviour in Twitter Data”, Amrita Mangaonkar, Department of Computer Sciences Indianapolis, Indiana, August 2017..
- [5]“Sarcasm Detection of Tweets: A comparative Study”,Tanya Jain, Nilesh Agrawal,Garima Goyal1,Proceedings of 2017 Tenth International Conference on Contemporary Computing (IC3), Noida, INDIA,10-12 August 2017.

Mobile Tool for analysis of Events,stocks and Management System

Rahul Nair, Justin James, Rishab Koul , Mehmood Deshmukh, and Prof.Deepti Lawand

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

Abstract— The stock is a valuable asset for any business and also probably most susceptible to pilferage, damage, expiry, wastage or fraud. The objective of stock verification is to prove the existence, accuracy, ownership rights and ensure the realizable value of the items in Company's inventory. Since the inventory has many movements on business days, the process of routine physical verification become a difficult task for any organization. It needs proper planning, resource mobilization, and expertise. Therefore combining the processes of stock management and event management using a single mobile tool is done. In this project we are implementing a software to manage an event where we will provide all the equipment list based on the requirements given by the client. We can also hire the items from a third party seller and make it available to the client. The main aim of this project is to reduce the communication gap between the client and the responsible persons of the company and hence reducing the paperwork and tedious tasks. All these processing will be done with the help of an mobile application. We are going to use various layouts such as linear, relative,constraints etc. All our confidential data and credentials will be stored in a very secured database i.e Firebase. The main module in this project will be that the processing of the customer requirements according to the nearest warehouse available hence provided fastest supply of the required items. Floyd Warshall algorithm will be used to track the available stock in the warehouses and a navigation system will be created for the same.

Keywords— Floyd Warshall, Inventory, Stocks, Event, Navigation, Database.

1. Introduction

Smartphone is a common computational device that possessed by the most of people nowadays, which is the inspiration to create an application that its information can be easily reached anywhere, any time. In addition, it would be difficult to manage all event registration manually, because it will take a long time for a long queue of customers to sign their name at the registration table, also a lot of document to handle. Furthermore, people nowadays prefer convenience for their life. In other words, it is harder for users to open the website then click on an application in their smartphones. The interior of storage management can be considered to be a kind of layered management and its exterior, together with related entities, such as supplier and customer, etc, forms dynamic network system.

2. Literature Survey

A. Development of Inventory management System

It is mainly responsible for the management of the domain, representing manufacturer to interact with material supplier Agent, making bidding plan after accepting a task, receiving a bid before deadline, selecting a suitable bidder according to improved contract net protocol, sending transaction information to material supplier Agent, negotiating and communicating with production manager Agent to determine order quantity and cycle, handling the material demand information of purchasing Agent and carrying out inventory control with storage management Agent. It is mainly responsible for the warehouse-in and warehouse-out of various materials and corresponding cost management in storage, reporting storage management work to storage manager. Agent in time and receiving the feedback information of inventory Agent.^[1]

B. Smart Mobile-based Notification System

We propose a convenient and user-friendly disaster and emergency management system that includes a server, a mobile application synchronized with a smart watch, and an accompanying website designed for disaster relief authorities, such as the concerned governmental agencies. Our proposed web portal enables governmental agencies to alert users immediately of emergencies as well as maintain the credibility of the alerts. That is because the web portal allows the agencies to view alerts about possible emergencies from users for further investigation about their seriousness. It also enables the notification of all users within the affected radius of an emergency incident. Once a crisis occurs, the concerned disaster management authority, such as the local police, can locate the affected region on the map using the proposed accompanying web interface for governmental agencies.^[3]

C. Event Management System

Online event management system is an online event management system software project that serves the functionality of an event manager. The system allow only registered user login and new user are allowed to register on the application .This proposed to be a web application. The project provides most of the basic functionality required for an event type e.g. [marriage, Dance Show birthday party, etc.], the system then allows the user to select date and time of event, place and the event equipment. All the data is logged in the database and the user is given a receipt number for hisbooking. The data is then send to administrator (website owner) and they may interact with the client as per his requirement.^[2]

2.1 Summary of Related Work

The overview of comparison of different parameters are given in Table

Sr. No.	Title	Advantages	Disadvantages
1	Yang Fan, Development of Inventory management System 978-1-4244-5265-1/10/ ©2010 IEEE	Full inventory management	No sales tracking
2	Phanuphong Hathaiwichian, Laps Siriwittayacharoen , Android Application for Event Management 978-1-4799-5573-2/14/ ©2014 IEEE	Live ticket management	No inventory management
3	Mohammed Ghazal, Samr Ali, Marah Al Halabi, Nada Ali, and Yasmina Al Khalil, Smart Mobile-based Emergency Management and Notification System 978-1-5090-3946-3/16 ©2016 IEEE	Direct free sms service	-

Table 1 Summary of literature survey

3. Proposed Work

This mobile tool is used as both storage management and also as a navigation device .This solves the problem of having different application for these two different functions. This tool helps in reduction of paper-pen work and also helps to reduce the to and fro communication by the salesman on which warehouse to go and which warehouse has the items that the customer requires and also we are using floyd warshall algorithm to find the shortest distance to the warehouse which reduces the time taken to deliver the product by fair amount.

3.1 System Architecture

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

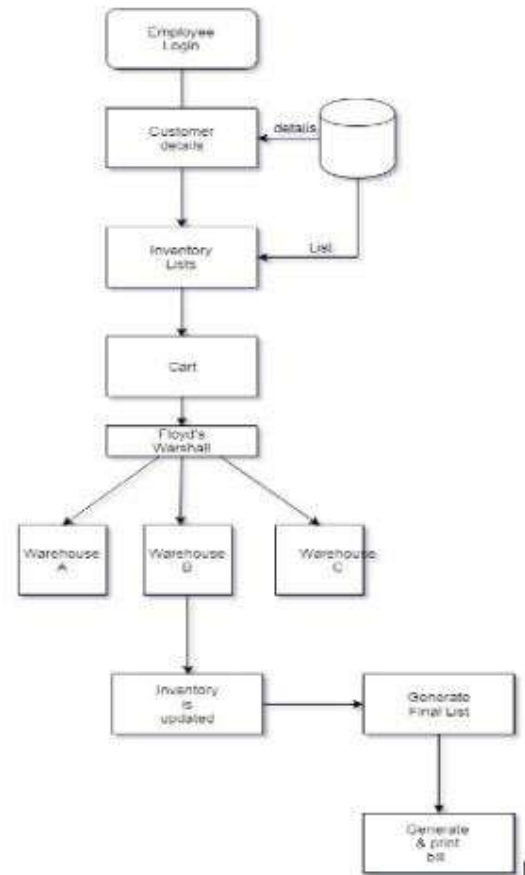


Fig. 1 Proposed system architecture

A. Input Stage Description: The first part of the system defines the login of the employee which is used to take the customer details at a particular place. As soon the details of the customer are taken as the input to the employee the following details are stored in the database. Further The requirement which is needed by the customer is taken the next input. The following process is carried through a smartphone where the customer will be provided with a checklist of the items. Therefore all the requirement is created as list of items and is stored in the database. The details which are stored in the database is used at the time of bill preparation.

B. Floyd Warshall Algorithm : Our system will be implemented using Floyd Warshall Algorithm. This is an algorithm for finding shortest paths in a weighted graph with positive or negative edge weights (but with no negative cycles). A single execution of the algorithm will

find the lengths (summed weights) of shortest paths between *all* pairs of vertices. Although it does not return details of the paths themselves, it is possible to reconstruct the paths with simple modifications to the algorithm. Versions of the algorithm can also be used for finding the transitive closure of a relation R $\{\displaystyle R\}$, or (in connection with the Schulze voting system) widest paths between all pairs of vertices in a weighted graph.

C. Overall Description: The second part receives a list of items required by the customer for the event. The items finalized in the cart is searched in the inventory of the particular company in realtime. Hence reducing the communication gap between the employee and the higher authorities. There are multiple warehouses in our system and each of them are tracked through Google API. A navigation device is associated with the employee and whenever he visits the customer and takes input of the inventory the shortest distance is calculated with the help of Floyd Warshall. The shortest distance when calculated is assigned to the inventory of the nearest warehouse from which the inventory is to be decremented.

E. Output Block Description: After the process is finished a final bill is generated for the customer in a sorted manner. To avoid fraud disturbance everything is done in real time. The finalized bill will contain the details of the items used and the discount given to the customer after communicating with the higher authorities and hence a soft-copy of the bill is sent to the customer's email id in real time.

3 Requirement Analysis

The implementation detail is given in this section.

3.1 Software

Software used for implementing the proposed system will be Android Studio. Android Studio is the official integrated development environment (IDE) for Google's Android operating system, built on JetBrains' IntelliJ IDEA software and designed specifically for Android development. It is available for download on Windows, macOS and Linux based operating systems. It is a replacement for the Eclipse Android Development Tools (ADT) as the primary IDE for native Android application development.

3.2 Hardware

The first step of this algorithm is to convert the input address and navigate it to the nearest warehouse. This is done by making the coordinate system equal to the entire pixels of the warehouses around the given customer address. By doing so, the warehouse inside any given map can be calculated using the coordinate values. The

application is the real time software (mobile application) and therefore there is no use of specific hardware device.

ACKNOWLEDGMENT

It is our privilege to express our sincerest regards to our supervisor **Prof. Deepti Lawand** for the valuable inputs, able guidance, encouragement, whole-hearted cooperation and constructive criticism throughout the duration of this work. We deeply express our sincere thanks our Head of the Department **Dr. Madhumita Chatterjee** and our Principal **Dr. Sandeep M. Joshi** for encouraging and allowing us to presenting this work. We extend our sincere appreciation to all our Professors from Pillai College of Engineering for their valuable insight and tips during the designing of the project. Their contributions have been valuable to us in so many ways.

REFERENCES

- [1] *Yang Fan*, Development of Inventory management System 978-1-4244-5265-1/10/ ©2010 IEEE.
- [2] *Phanuphong Hathaiwichian, Lapas Siri Wittayacharoen*, Android Application for Event Management 978-1-4799-5573-2/14/ ©2014 IEEE.
- [3] *Mohammed Ghazal, Samr Ali, Marah Al Halabi, Nada Ali, and Yasmina Al Khalil*, Smart Mobile-based Emergency Management and Notification System 978-1-5090-3946-3/16 ©2016 IEEE.
- [4]. *Punam Khobragade, Roshni Selokar, Rina Maraskolhe Prof. Manjusha Talmale (2018)*, Research paper on Inventory management system, International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056.
- [5] *M. O Yinyeh, S. Alhassan*, Stock Management System Software for Public Universities in Ghana (IMSSPUG), International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 2, Issue 8, August 2013, ISSN: 2278 – 1323.
- [6]. *Amir Saleem, Davood Ahmed Bhat, Mr. Omar Farooq Khan*, Review Paper on an Event Management System, International Journal of Computer Science and Mobile Computing, IJCSMC, Vol. 6, Issue. 7, July 2017, pg.40 – 43.
- [7]. *Vinay Mishra, Madhuri Dubey, Priya Banerjee, Ajvita Jumle, Pallavi Raipureand, Pooja Wankhede*, Event Management System International Journal of Trend in Research and Development, Volume 3(6), ISSN: 2394-93.