The 1st International Conference on Innovations in Computer Science & Software Engineering

NEDUET iCONICS 2016 KARACHI

December 15-16, 2016

NED University of Engineering & Technology, Karachi Pakistan

Editor: Dr. Muhammad Mubashir Khan

Assistant Editors: Ms. Uzma Zehra, Ms. Sidra

In cooperation with the Higher Education Commission Pakistan

A Publication of The Department of Computer Science & Software Engineering, NED University of Engineering & Technology, Karachi, 75270 Pakistan

PREFACE

This conference booklet contains the abstracts of contributions submitted in the 1st International Conference on Innovations in Computer Science & Software Engineering (ICONICS-2016) held at Department of Computer Science & Software Engineering on December 15-16 2017. The purpose of the conference is to provide a platform for researchers both national and international to exchange novel and contemporary ideas in the emerging fields of computing and to provide a forum for researchers from academia and industry to meet and share cutting-edge development in the field of Computer Science & Software Engineering.

The conference looks to unite a wide range of global specialists to make a helpful situation for coordinated effort and learning exchange. The conference requests huge commitments in every real field of Computer Science and Software Engineering from theoretical and useful viewpoints. Authors are invited to contribute to the conference by submitting articles that illustrate research results, surveying works and industrial experiences that describe significant advances in the field of Computer Science and Software Engineering. It presents new findings on theoretical or practical aspects of computer science and software engineering, share experiences on combining new technologies into products and applications, discuss the work on applying computing and software engineering to real-life problems, describe their development and operations of challenging computing related systems, and highlight open research problems.

All pre-conference arrangements were highly successful. We record our gratitude to our kindred individuals from the dedicated organizing committee and distinguished technical program committee for their work in securing a considerable contribution of research articles. We additionally recognize the authors themselves, without their contribution we could not imagine such a successful conference.

ORGANISING COMMITTEE

Conference Convener:	Prof. Dr. Muhammad Tufail
Conference Co-Convener:	Prof. Dr. Najmi Ghani Haider
Conference Co-Convener:	Prof. Dr. Sohail Abdul Sattar
Conference Secretary:	Dr. Shariq Mahmood Khan
Finance Committee:	Dr. Sh. M. Wahabuddin Usmani (Chair)
	Mr. Kashif Mehboob Khan
Travel & Accommodation Committee:	Dr. Najeed Ahmed Khan (Chair)
	Mr.Waseemullah
	Mr. Umer Farooq
	Mr. Nadeem Ahmed Khan
Website Management Committee:	Dr. Shariq Mahmood Khan (Chair)
Abstract & Poster Committee:	Dr. Saman Hina (Chair)
	Mrs. Nazish Irfan
Reception Committee:	Ms. Saba Izhar Haque (Chair)
	Ms. Maria Andaleeb Siddiqui
	Ms. Uzma Zehra
	Ms. Sidra
Publicity Committee:	Dr. Shehnila Zardari (Chair)
	Mr. Muhammad Imran Shaikh
	Ms. Uzma Zehra
Publications Committee:	Dr. Muhammad Mubashir Khan(Chair)
	Ms. Uzma Zehra
	Ms. Sidra
Registration Committee:	Engr. Raheela Asif (Chair)
	Ms. Sidra
	Mr. Nadeem Ahmed Khan
Sponsors Committee:	Engr. Amjad Ali (Chair)
	Ms. Noor Afshan Vasty
	Ms. Maria Andaleeb Siddiqui.
Catering Committee:	Mr. Mehboob Ahmed (Chair)
	Mr. Imdadullah
IT Support Committee:	Mr. Muhammad Ejaz Yunus (Chair)
	Ms. Noor Afshan Vasty
	Mrs. Nazish Irfan
	Mr. Muhammad Imran Shaikh
	Mr. Mehboob Ahmed

TECHNICAL PROGRAMME COMMITTEE

Dr. Liang-Ting Chen	University of Hawaii at Manoa, US
Prof. Dr. Junaid Ahmed Zubairi	The State University of New York, US
Prof. Dr. Agathe Merceron	Beuth University of Applied Sciences, Germany
Dr. Faraz Hasan	Massey University, NZ
Dr. Muhammad Atif Tahir	Northumbria University, UK
Dr. Junaid Arshad	University of West London, UK
Dr. Sakinah Ali Pitchay	Universiti Sains Islam Malaysia
Dr. Nasrullah Memon	University of Southern Denmark
Dr. Tafseer Ahmed Khan	DHA Suffa University, PK
Dr. Imran N. Junejo	University of Sharjah, UAE
Dr. Jawad Ahmed Shamsi	FAST National University, PK
Dr. Muazzam Siddiqui	King Abdulaziz University, KSA
Dr. Syed Raheel Hassan	Quaid-E-Awam University, PK
Dr. Mukhtaj Khan	Abdul Wali Khan University Mardan, PK
Dr. Miriam Paiola	Independent Researcher, France
Dr. Izaz ur Rehman	Abdul Wali Khan University Mardan
Dr. Mahamat Issa Hassan	University of King Faisal Chad
Dr. Naeem Ahmed Mahoto	Mehran University of Engineering and Technology PK
Dr. M. Ajmal Sawand	Institute of Business Administration Sukkur, PK
Dr. Imran Sarwar Bajwa	The Islamia University of Bahawalpur, PK
Dr. Safeeullah Soomro	AMA International University Bahrain
Dr. Shahzad Ahmed Memon	University of Sindh, PK
Dr. Shahabuddin Muhammad	Prince Mohammad Bin Fahd University, KSA
Dr. Quratul-ain Mahesar	SZABIST Hyderabad, PK
Dr. Ghufran Ahmed	COMSATS Islamabad Campus, PK
Dr. Esra Alzaghol	University of Jordon
Dr. Syed Abbas Ali	NED University of Engineering & Technology, PK
Dr. Hassan Ali Baig	Institute of Geo-Information and Earth Observation, PK
Dr. Muhammad Asad Arfeen	NED University Karachi, PK
Dr. Hussien Al-Hmood	Thi-Qar University Iraq
Dr. Thomas Andzi-Quainoo Tawiah	University of Education Winneba, Ghana
Dr. Shukat Wasi	Mohammad Ali Jinnah University, PK
Dr. Syed Tahir Qasim	FAST National University, PK

SPONSORS


HEC
HIGHER EDUCATION COMMISSION


AbacusConsulting
Your Transformation Partner


JBS
Works for you


mare
CONSULTING


Education
USA
A U.S. Department of State Network


The United States
Educational Foundation
in Pakistan


SUZUKI

TABLE OF CONTENTS

ICONICS 2016 KEYNOTE

IoT (Internet of Things) and Its Applications

Junaid Ahmed Zubairi
The State University of New York, USA

ABSTRACT

IoT (Internet of Things) is the extension of networking, computing and data generation to everyday objects and sensors in automated and autonomous ways. IoT envisages a world in which devices take independent decisions for reaching a predefined goal with minimal human intervention. It is estimated that over 50 billion devices would be connected by 2020, each device capable of sensing, analyzing and sending data wirelessly. Some key components of IoT include sensors, aggregators, actors and e-utilities including analysis and decision software. IoT applications are in the broad fields of transportation, e-health, warfare, utilities, industrial automation and municipal services. IoT can be divided into broad categories of Industrial and Consumer platforms. Both of these categories have plenty of exciting developments taking place. In this talk, I will present an introduction of IoT, its challenges, issues and current and potential applications.

KEYNOTE SPEAKER'S BIOGRAPHY

Prof. Dr. Junaid Ahmed Zubairi received his BE (Electrical Engineering) from NED University of Engineering, Pakistan and MS and Ph.D. (Computer Engineering) from Syracuse University, USA in 1991. He worked in Sir Syed University Pakistan and Intl' Islamic University Malaysia before joining State University of New York at Fredonia, USA in 1999 where currently he is a Professor in the Department of Computer and Information Sciences. Dr. Zubairi is a recipient of many grants and awards including Malaysian Government IRPA research award, NSF MACS grant and multiple SUNY scholarly incentive awards. His research interests include network traffic engineering, network protocols and applications of networks. He has edited two books on network applications and security and published over fifty peer-reviewed chapters, journal articles and conference proceedings papers. He can be reached at zubairi@fredonia.edu.

Selfish Routing vs Cooperative Routing for congestion control in Wireless Networks

Nasreen Badruddin
Universiti Teknologi PETRONAS, Malaysia

ABSTRACT

Most of the practical routing techniques used are based on selfish/greedy algorithms that optimize an objective function from the viewpoint of each individual. Several literature has reported that selfish and uncoordinated behaviours of nodes can lead to overall degradation in performance of the network. Despite their inefficiencies, selfish routing algorithms are still implemented in practical networks, due to the lack of centralized regulation and coordination between the nodes in the network.

In cooperative routing, nodes work together by sharing information so that a route that optimizes a global network-wide objective function is achieved. There has been some work done in cooperative routing in communication networks, mainly with the objective of minimizing power consumption and network lifetime maximization. In this talk, we will present the research work done cooperative routing for the purpose of congestion control when directing traffic. Through information sharing between nodes, a routing scheme can perform better in achieving congestion avoidance, detection and alleviation.

KEYNOTE SPEAKER'S BIOGRAPHY

Dr. Nasreen Badruddin graduated with a first class honours B.Eng. degree in Electronic Engineering from RMIT University, Australia, in 2000. She later received M.Sc. in Electrical & Computer Engineering from Carnegie-Mellon University, USA in 2002, and a Ph.D. in Electrical & Electronic Engineering from the University of Melbourne, Australia, in 2011. Her research interests are primarily in the area of wireless communications and networks as well as biomedical engineering, particularly in neuro-signal processing and wireless body area networks (WBAN). Nasreen was a recipient of a PETRONAS overseas scholarship for her undergraduate studies and an Endeavour Postgraduate Award from the Australian government for her Ph.D. studies. She is currently an Associate Professor and Deputy Head of the Department of Electrical and Electronic Engineering, at Universiti Teknologi PETRONAS. She is attached to the Biomedical Technology research group as well as the Centre of Intelligent Signal and Imaging Research (CISIR).

Multilabel Machine Learning Algorithms for Text, Image, and Video Classification

Muhammad Atif Tahir
Northumbria University, UK

ABSTRACT

Multi-label classification is a challenging research problem due to the fact that each example may belong to a varying number of classes. This is different from standard multi-class classification system, where only a single label is predicted for each sample from a set of distinct labels. This problem can be further aggravated by high dimensionality and complex correlation among labels. In this talk, I will discuss state-of-the-art multi-label classification with applications in image / video retrieval and text classification including my contributions in this emerging field.

KEYNOTE SPEAKER'S BIOGRAPHY

Dr. Muhammad Atif Tahir received his PhD from School of Computer Science & Engineering at Queens University, Belfast, UK, MSc in Computer Engineering from King Fahd University, Dhahran, KSA, and BE in Computer Systems Engineering from NED University of Engineering, and Technology, Karachi, Pakistan. He is currently working as Senior Lecturer at School of Computer Science, Northumbria University, United Kingdom. His main research is in Machine Learning & Combinatorial Optimization Techniques with applications in image / video retrieval, cancer classification, surface inspection, bioinformatics, multi-label classification, and face recognition. He has authored and co-authored more than 60 publications in top quality journals including IEEE Transactions on Pattern Analysis and Machine Intelligence, Pattern Recognition, Journal of Machine Learning Research, IEEE Transactions on Multimedia. From January 2017, he will join Karachi Campus of National University of Computer & Emerging Sciences (FAST-NU) as Professor in Computer Science.

Content Optimization for Neuro feedback based Treatment of Stress and Depression

Syed Saad Azhar Ali
Universiti Teknologi PETRONAS, Malaysia

ABSTRACT

People of different age, sex, status and job conditions are suffering from different mental states like stress, depression, memory loss etc. and disorders like epilepsy, autism, etc. These people are part of our society and we cannot neglect their presence. Contrary to conventional medicinal treatment, that may cause injurious side effects, a noninvasive approach employing neurofeedback is used recently by the clinicians and psychiatrists for diagnosis and treatment. Neurofeedback treatment has shown promising results for the treatment of such conditions. There are different modes and contents used for neurofeedback like music, videos, images, games, reading, words counting, addition, the color of words, etc. Clinicians are still exploring the best ways to employ neurofeedback. However, it is still premature in the method that can decide which mode or content will result in better treatment. It is also reported that due to non-efficacy of the treatment the subjects have withdrawn treatment, resulting in loss of time, cost and most importantly the impact could be drastic especially for depression patients. This research focuses on investigating the right mode and content for neurofeedback based treatment. The investigation will be based on selecting the optimal content to be used during neurofeedback treatment depending upon subject history, profile, personality, mental state, etc. The expected outcome of this investigation is a systematic approach that will assist in choosing the optimal content that will ensure treatment efficacy right from the early stages. This research will make sure that the patients will continue having treatment and societal burden is eventually reduced.

KEYNOTE SPEAKER'S BIOGRAPHY

Dr. Syed Saad Azhar Ali obtained his BE from NED University in Electrical Engineering. He secured his Masters and PhD from King Fahd University of Petroleum & Minerals in the area of Nonlinear Control. He has been associated with Air University and Iqra University prior to being engaged with the Center of Intelligent Signal & Imaging Research at Universiti Teknologi PETRONAS. His research focus has been on Intelligent Control, Signal Processing, and Underwater robotics with the emphasis on image enhancement and 3D scene reconstruction. He is leading the Visual Surveillance and Monitoring Group. Recently he has been involved in neuro-signal processing as well. He is the author of over 60 peer-reviewed publications including 4 books/chapters. Dr. Ali is an IEEE Senior Member. He is the PI for several Funded Research Projects.

How Big Data is changing the world?

Muhammad Adnan
Marks and Spencer, UK

ABSTRACT

The amount of data collected since the dawn of humanity until 2003 was equivalent to the volume we now produce every two days. This data comes not only from social media posts, documents, emails, business applications, machine log data, media but increasingly from sensors on objects from lamp-posts to skyscrapers. Big Data is increasingly being used by prominent companies to outpace the competition. Be it established companies or start-ups, they are embracing data-focused strategies. Companies have infinite amounts of data that can inform what customers are actually looking for. Data informs what is built tomorrow and how the current products can be improved. For example, 852 Marks and Spencer stores around the UK generate more than 5 million transactions daily. This big data is not only used to predict the sales of more than 30,000 products in coming months but also to predict what each customer will likely to buy in the future. This keynote will outline the advances in the field of Big Data Analytics with a special focus on its application in business.

KEYNOTE SPEAKER'S BIOGRAPHY

Dr. Muhammad Adnan completed a Bachelor degree in Computer Science & IT from N.E.D. University of Engineering & Technology in 2005. Later, he received an M.Sc. in Software Engineering from Queen Mary, University of London in 2007. He joined University College London as a Data Science Researcher in 2007, before undertaking a part-time Ph.D. in GIScience (2008-2011). After completing the Ph.D., he was appointed as a Senior Research Associate by Consumer Data Research Center run by University College London. He left UCL in 2015 and joined Royal Dutch Shell PLC as a Data Scientist. Currently, he works for Marks and Spencer PLC, a major British retailer, as a Data Scientist and helps shape the big data analytics work within Marks and Spencer's foods business. Muhammad Adnan has research interests in Big Data Analytics, Information Visualization, Visual Analytics and their applications to a variety of uses including demographics, socio-economic analysis, supply chain, and demand forecast etc. Many of his visualizations and research work outputs have been featured in a number of media outlets including the Guardian, the Evening Standard and the Daily Mail.

ICONICS 2016 PAPERS

The 1st International Conference on Innovations in Computer Science & Software Engineering, (ICONICS 2016), December 15-16, 2016

A Comparative Study for Software Cost Estimation Using COCOMO-II and Walston-Felix models

Amanullah
Dept. Computing and Technology Absayn University Peshawar
scholar.amankhan@gmail.com

Abdus Salam
Dept. Computing and Technology Abasyn University Peshawar
dr.salam@abasyn.edu.pk

Abdullah Khan
Institute of Business and Management Science University of Agriculture Peshawar
abullahdirvi@gmail.com

Samad Baseer
Dept. Computer System Engineering UET Peshawar
drsamadbaseer@uetpeshawr.edu.pk

***Abstract*— Software cost estimation is an important area in software project management life cycle. Accurate estimation of the effort in term of person month and development time is required before startup of a project. There are many software cost estimation models such as algorithmic and non-algorithmic. This paper presents the comparison between two models that is Constructive Cost Model II (COCOMO-II) and Walston-Felix models. The experiment is conducted on NASA and Turkish industry software projects. From the simulation results it shows that the COCOCMO-II has the better result than the Walston-Felix in term of Mean Magnitude of Relative Error (MMRE) for both dataset.**

Keywords—COCOMO-II, Walston-Felix, MMRE, NASA project, Turkish Industry software projects

I. INTRODUCTION

Cost estimation is one of the essential part of software development life cycle (SDLC). This is also important for managing project cost and time. Software project management consist of two components: i) Project Planning ii) Project Monitoring and Control. Planning is estimating the activities that must be performed before starting development task. Once software project task is initiated, it is responsibility of project coordinator to monitor these task and conform that the high quality software is produced with minimum cost and within a specified time. One of the crucial elements of software cost estimation is predict the amount of time required to develop software product. The time is measured in term of Person Month (PM) which changed later on into currency cost. The basic input for cost and efforts estimation or prediction model is the Source Line of Code (SLOC) in term of Kilo Source Line of Code (KSLOC) and set of cost drivers [1]. There are two type of cost estimation models: i) algorithmic ii) non-algorithmic. The most prevalently used non-algorithmic models are Expert Judgment (EJ) methods, Estimating by Analogy, Top-down estimating method, Bottom up estimating method, Parkinson's law and Price-to win. The algorithmic models based on mathematical equation to perform estimation. Some of popular algorithmic methods are COCOMO (Constructive Cost model), Walston-Felix Model (WF), Baily-Basil model, and Halstead model, Doty (for KLOC > 9), Putnam Model, Function Point Analysis, SEL Model and COCOMO-II Model. In this works, the performance of Constructive Cost Model (COCOMO-II) and Walston-Felix (WF) will be compared to calculate cost estimates more accurately. The experiment has been conducted on two historical data set that is NASA-93 and Turkish Industry software projects. The rest of paper is organized as follow: Section II discuss related works, while section III illuminates COCOMO-II model. Walston- Filex model is explain in Section IV, furthermore section V, VI, VII, and VIII show the results analysis, evaluation criteria, dataset description conclusion and future work.

II. RELATED WORKS

In this part of the paper we discuss some pervious papers about software cost estimation techniques. Software cost estimation is an important aspect in software development [2]. Many software cost estimation models and methods have been proposed over the last ten years [3]. Accurate effort and cost estimation of the software projects continues to be a serious problem for software project managers[4]. According to T.N Sharma in Putnam model the time is very dominating factor while in COCOMO-II model depend up on cost driver and scale factor [5], and in [6] the comprehensive analysis of software cost estimation were discussed. In [7] a novel techniques the expert judgment is use to predict the cost of software projects. In this technique, the expert opinion can be described as judgment of individual expert or group of experts with respect to a particular subject. Furthermore in [8] Expert judgment explained which consist of two approaches, namely, top down and bottom up approaches. Another common technique for cost estimation is prediction by analogy, it is based on clime that project efforts to be estimated, will probably behave like efforts of similar previous projects [9]. In paper [15] the author proposed two methods to support resolving software effort estimation model problems. These techniques are Particle Swarm Optimization (PSO) and Fuzzy Logic (FL). Existing techniques only used PSO to optimize coefficient of the COCOMO-II and compare it with Fuzzy Logic (FL) method. While my study compares Walston-Felix and COCOMO-II for NASA-93 and Turkish industry software

projects. From the simulation analysis it is evidence that COCOMO-II has more accurate result than Walston-Felix in terms of Mean Magnitude of Relative Error (MMRE).

III. COCOMO-II MODEL

COCOMO-II (Constructive Cost Model II) [10], which was proposed by Barry Boehm in 1995, is a technique used for estimation of project cost, effort and calendar time. In COCOMO-II, the effort is represented in term of Man Month (MM). A man month is the amount of period one person spends to work on software development project for one month [11]. The COCOMO-II formula of effort for software projects is given in equation 1.

$$MM = A.\ size^{e}.\prod_{i=1}^{17} EMi \qquad (1)$$

Where A is constant, having value of 2.94. Size is represented in Kilo Source Line of Code (KSLOC). There are 17 cost drivers called Effort Multiplier (EM) given in Table I. The exponent E in equation 1 is a collection of five scale factor which is calculated in equation 2.

$$E=B+0.01\cdot\sum_{i=1}^{5} SF \qquad (2)$$

Where B is constant having value of 0.91. The Scaling Factors (SF) rating ranges from very low to extra high. Which are given in Table II. The software industry requires to calculate the amount of time that is Time Development (TDEV) of software project. Which is computed as

$$TDVE(t) = C \times MM^{F} \qquad (3)$$

Where

$$F = D+0.2\times 0.01\times \sum_{i=1}^{5} SF \qquad (4)$$

The value of C = 3.67 and D =0.28 are constant values according to COCOMO-II model.

TABL I: COST DRIVERS FOR COCOMO-II

Drivers	Very Low	Low	Nominal	High	Very High	Extra High
RELY	0.82	0.92	1.00	1.10	1.26	--
DATA	--	0.90	1.00	1.14	1.28	--
CPLX	0.73	0.87	1.00	1.17	1.34	1.74
RUSE	--	0.95	1.00	1.07	1.15	1.24
DOCU	0.81	0.91	1.00	1.11	1.23	--
TIME	--	--	1.00	1.11	1.29	1.63
STOR	--	--	1.00	1.05	1.17	1.46
PVOL	--	0.87	1.00	1.15	1.30	--
ACAP	1.42	1.19	1.00	1.85	0.71	--
PCAP	1.34	1.15	1.00	0.88	0.76	--
PCON	1.29	1.12	1.00	0.88	0.81	--
APLEX	1.22	1.10	1.00	0.90	0.81	--
APLEX	1.19	1.09	1.00	0.91	0.85	--
LTEX	1.20	1.09	1.00	0.91	0.84	--
TOOL	1.17	1.09	1.00	0.90	0.78	--
SITE	1.22	1.09	1.00	0.93	0.86	0.80
SCED	1.43	1.14	1.00	1.00	1.00	--

TABLE II: SCALE FACTOR FOR COCOMO-II

Scale Factors	Very Low	Low	Nominal	High	Very High	Extra High
PREC	6.20	4.96	3.72	2.48	1.24	0.00
FLEX	5.07	4.05	3.04	2.03	1.01	0.00
RESL	7.07	5.65	4.24	2.83	1.41	0.00
TEAM	5.48	4.38	3.29	2.19	1.10	0.00
PMAT	7.80	6.24	4.68	3.12	1.56	0.00

IV. WALSTON-FELIX MODEL

Walston-Felix model was proposed by Walston and Felix in 1977. Their effort model is based upon various phases of software development environment such as user database of 60 software projects collected in IBM's Federal System division [12].

According to Walston and Felix effort is calculated by

$$Effort = 5.2(L)^{0.91} \qquad (5)$$

$$Duration\ D = 4.1(L)^{0.36} \qquad (6)$$

Where 5.2, 0.91 4.1 1nd 0.36 are constant. L represents the KSLOC (Kilo Source Line of Code).

V. EVALUATION CRITERIA

The performance measures deliberated in our work were Mean Magnitude of Relative Error (MMRE). In order to evaluate these techniques and the appropriate accuracy of COCOMO-II in comparison to that of Walston-Felix technique, we used the most common evaluation measures in the field of software engineering. These are Magnitude of Relative Error (MRE) and Mean Magnitude of Relative Error (MMRE).

Magnitude of Relative Error (MRE) calculates the absolute percentage of errors among actual and predicted effort for each reference software project.

$$MRE = \left|\frac{\text{Predicted Value} - \text{Actual Value}}{\text{Actual Value}}\right| \qquad (7)$$

Mean Magnitude of Relative Error (MMRE) computes the average of Magnitude of Relative Error (MRE) of software projects.

$$MMRE = \frac{1}{N}\sum_{i=1}^{n} \frac{\text{Predicted Value} - \text{Actual Value}}{\text{Actual Value}} \qquad (8)$$

Where N is total numbers of software projects. MMRE represents the mean magnitude of relative error.

VI. DATA SET DESCRPTION

The data taken from NASA-93 and Turkish Industry software projects found in [13, 14]. These projects consist of (KLOC) thousand source line of code, actual effort and calendar time. The cost drivers and scale factors rating from very low to extra high are also present in the data set software projects. We selected 15 software projects from NASA-93 randomly and overall15 projects form Turkish Industry to meet the comparison level of both projects.

VII. RESULT ANANYSIS

The two models were evaluated using NASA and Turkish industry datasets and their results are shown in Table III, IV, V and VI. The NASA dataset consist of 93 software projects from which 15 software projects were selected for simulation. While the Turkish Industry dataset contains 15 software projects. From the result analysis it shows that the mean magnitude of relative error of COCOMO-II is 0.388, for NASA dataset while 6.278 for Turkish Industry dataset. Similarly the mean magnitude of relative error of Walston-Felix are 0.603, 8.928 of NASA and Turkish Industry software projects respectively.

The results of both case studies also show that the entire data set, MMRE of COCOMO-II is less than that of Walston-Felix techniques. Furthermore, the performance of both COCOMO-II and Walston-Felix are explained in detail for NASA and Turkish industry software projects.

TABLE III: COMPARISON OF COCOMO-II AND ACTUAL EFFORT USING NASA-93 DATA SET.

Project No	KLOC	Actual Effort	COCOMOII	MRE Effort of COCOMOII
1	32.50	60.00	100.4	0.6733
2	19.70	60.00	65.30	0.0888
3	12.80	62.00	64.10	0.0338
4	48.50	239.0	218.8	0.0845
5	15.40	70.00	68.90	0.0294
6	16	114	68.50	0.399
7	177.9	1248	793.9	0.572
8	90	162	166	0.024
9	162	756	662.6	0.140
10	32	1350	422.2	0.687
11	24	430	93.5	0.783
12	50	1924.5	526.8	0.726
13	233	8211	1966.5	0.760
14	16.30	480	163.4	0.659
15	3.00	38.00	31.7	0.165

Table III explains the simulation results of COCOMO-II for NASA software project. In Table III there are 15 software projects which consist of Kilo Lion of Code (KLOC), actual effort, predicted effort that is COCOMO-II and magnitude relative error. The magnitude of relative error is the difference between predicted values minus actual values divide by actual values. In these fifteen projects COCOMO-II is best for project number eight because it has lowest MRE as compared to all these projects. While Figure 1 give the graphically representation of the result.

TABLE IV: COMPARISON OF WALSTON-FELIX AND ACTUAL EFFORT USING NASA-93 DATA SET.

Project No	KLOC	Actual Effort	Walston-Filex	MRE Effort of Walston-Felix
1	32.50	60.00	123.54	1.059
2	19.70	60.00	78.337	0.305
3	12.80	62.00	52.910	0.146
4	48.50	239.0	177.839	0.255
5	15.40	70.00	62.610	0.105
6	16	114	64.826	0.431
7	177.9	1248	580.314	0.535
8	90	162	312.15	0.926
9	162	756	532.91	0.295
10	32	1350	121.811	0.909
11	24	430	93.755	0.782
12	50	1924.5	182.837	0.904
13	233	8211	741.818	0.909
14	16.30	480	65.931	0.863
15	3.00	38.00	14.131	0.628

Similarly table IV shows the comparison between actual effort and Walston-Felix model in term of MRE for NASA fifteen software projects. The project five has lowest MRE value 0.105.Therefore Walston-Felix is best for project number five. Figure 2 shows the NASA data sets comparison for Walston-Felix model.

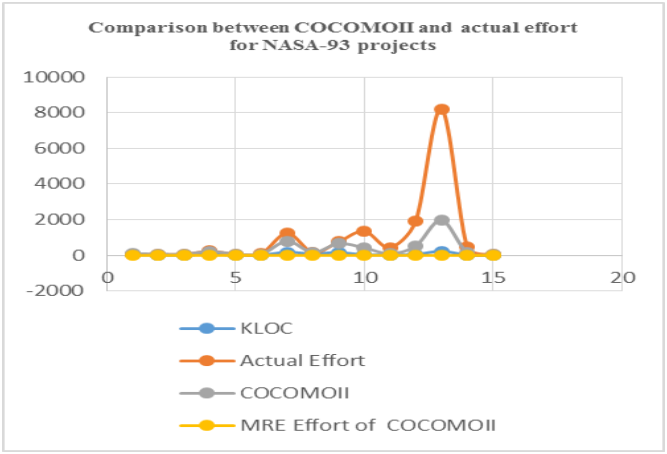


Fig 1: NASA data sets comparison on bases of COCOMO II

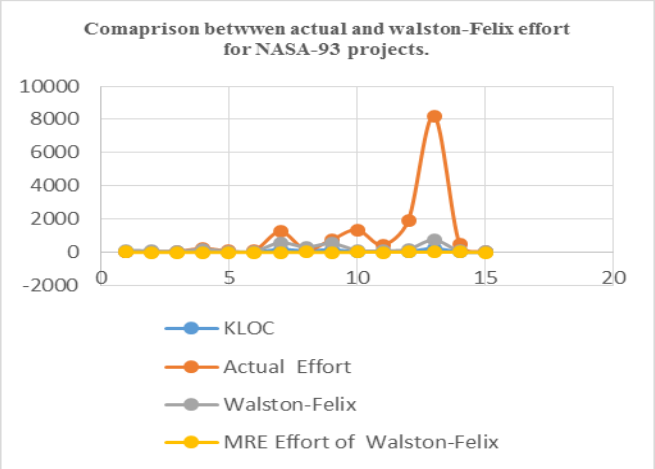


Fig 2: NASA data sets comparison on bases of Walston-Felix

TABLE V : MMRE of COCOMO II and WALSTON-FILEX For NASA-93 DATASET

Mean Magnitude of Relative Error (MMRE) of Effort	
COCOMOII	0.388
Walston-Felix	0.603

Table V represents the comparison in term of MMRE (Mean Magnitude of Relative Error) of Walston-Felix and COCOMO-II techniques. On the bases of experiment the COCOMO-II is better technique for effort estimation as compared to Walston-Felix model. The figure 3 shows the graphically representation of results.

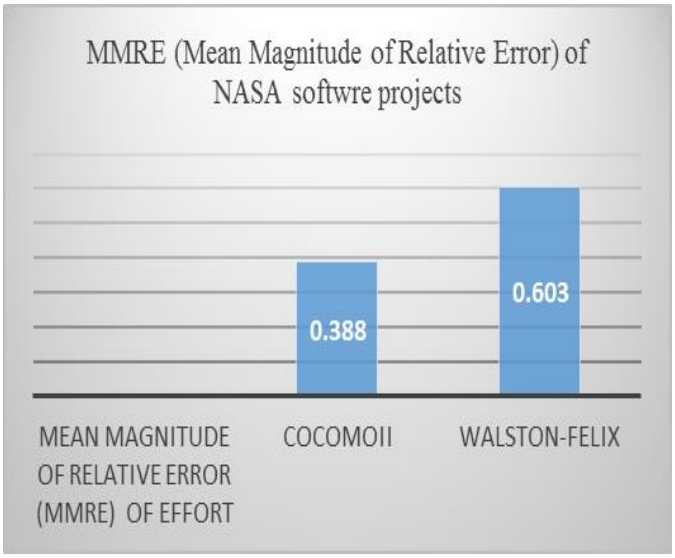


Fig 3: Comparisom between COCOMO-II and Walston-Felix in term of MMRE (Mean Magnitude of Relative Error) for NASA softwre projects.

TABLE VI COMPARISON BETWEEN COCOMO-II and ACTUAL EFFORT

Project No.	KLOC	Actual Effort	COCOMOII	MRE of COCOMOII
1	3.00	1.20	3.60	2.00
2	2.00	2.00	2.90	0.45
3	4.25	4.50	9.30	1.06
4	10.00	3.00	36.20	11.06
5	15.00	4.00	63.20	20.06
6	40.53	22.00	28.60	0.30
7	4.05	2.00	2.30	0.15
8	31.85	5.00	147.10	28.42
9	114.28	18.00	294.0	15.33
10	23.11	4.00	63.20	14.80
11	1.37	1.00	0.900	0.10
12	1.61	2.10	2.00	0.0476
13	196.6	638	722.70	0.132
14	51.80	185	176.50	0.045
15	64.10	332	256.7	0.220

The table VI shows the comparison of predicted effort i-e COCOMO-II efforts for Turkish Industry. In this dataset the project number 14 has lowest MRE. Therefore COCOMO-II is better technique for project 14 in this data sets. In figure 4 show the Turkish Industry data sets comparison for Walston-Felix model

Table VII denote the comparison between Walston – Felix model and actual efforts.In this dataset project 13 has lowest MRE. Therefore Walston-Felix model is better for project 13. The figure 5 represents the results graphically.

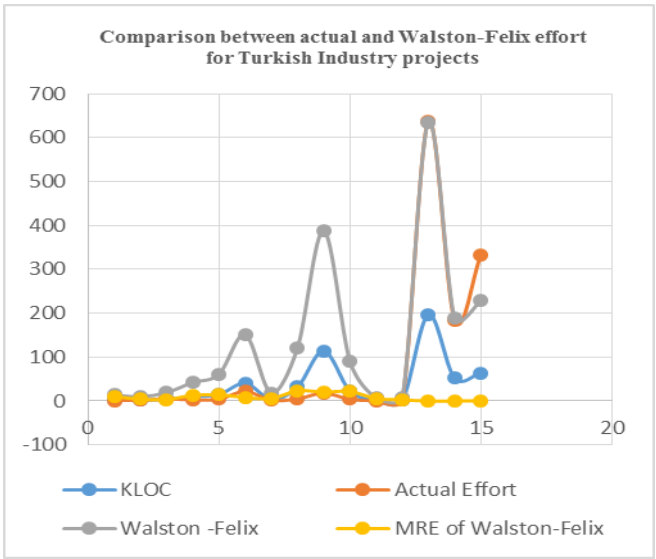


Fig 4.Turkish Industry comparison using the COCOMO-II

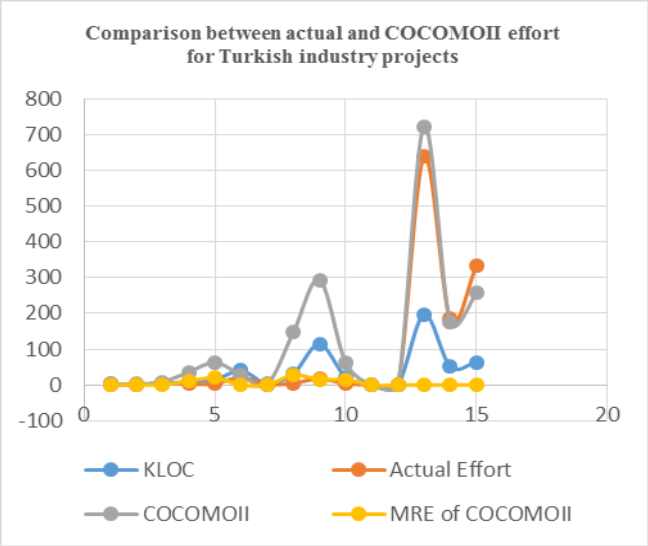


Fig 5: Turkish Industry data sets comparison using W.F

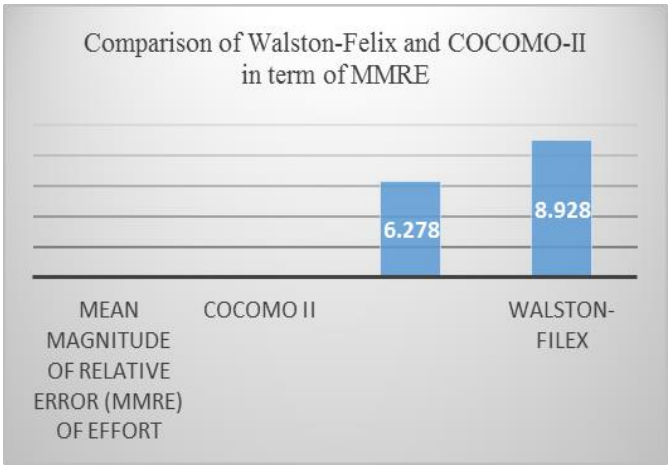


Fig 6: Comparison of Walston-Felix and COCOMO-II in term of MMRE for Turkish Industry dataset

TABLE VIII: MMRE of COCOMO II and WALSTON-FELIX FOR TURKISH INDUSTRY DATASET.

Mean Magnitude of Relative Error (MMRE) of Effort	
COCOMO II	6.278
Walston-Filex	8.928

Table VIII represents the Walston-Felix and COCOMO-II comparison in term MMRE.COCOMO-II is best technique for effort estimation as compared to Walston-Felix model. The figure 6 represents the results graphically.

VIII. CONCLUSION AND FUTURE WORK

Software Cost Estimation (SCE) is an important phase in software development life cycle (SDLC). Good estimation leads the project smoothly towards completion. In this work the two important software cost estimation models Walston-Felix and COCOMO-II were discussed. The COCOMO-II has lowest MMRE as compared to Walston-Felix technique, using NASA-93 and Turkish industry case studies by experimental analysis. Author conclusion which can be drawn from this work is that the COCOMO-II is better software cost estimation technique as compared to Walston-Filex Model for NASA-93 and Turkish industry software projects.

In future work we plan to compare Halstead model, Function Point Analysis (FPA) and COCOMO II models. This work can be further extended to compare the performance of new techniques for estimation by using Doty Model and COCOMO-II for various software projects.

ACKNOWLE DGMENT

The authors are thankful to Department of Computing and technology Abasyn University Peshawar Pakistan. My sincere gratitude goes to Dr. Abdus Salam Head of the Department of Computing and Technology Abasyn University Pakistan.

REFERENCES

[1] Ziauddin, S. K. Tipu, K. Zaman, and S. Zia, "Software Cost Estimation Using Soft Computing Techniques," *Advances in Information Technology and Management (AITM)*, vol. 2, pp. 233-238, 2012.

[2] A. S. Grewal, V. Gupta, and R. Kumar, "Emerging Estimation Techniques," *International Journal of Computer Applications*, vol. 59, 2012.

[3] S. Kumari and S. Pushkar, "Comparison and analysis of different software cost estimation methods," *IJACSA) International Journal of Advanced Computer Science and application*, vol. 4, 2013.

[4] K. Maxwell, L. Van Wassenhove, and S. Dutta, "Performance evaluation of general and company specific models in software development effort estimation," *Management Science*, vol. 45, pp. 787-803, 1999.

[5] T. Sharma, A. Bhardwaj, and A. Sharma, "A Comparative study of COCOMO II and Putnam models of Software Cost Estimation," *vol*, vol. 2, pp. 1-3, 2011.

[6] S. Kumari and S. Pushkar, "Performance analysis of the software cost estimation methods: a review," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 3, pp. 229-238, 2013.

[7] P. Faria and E. Miranda, "Expert Judgment in Software Estimation During the Bid Phase of a Project--An Exploratory Survey," in *Software Measurement and the 2012 Seventh International Conference on Software Process and Product Measurement (IWSM-MENSURA), 2012 Joint Conference of the 22nd International Workshop on*, 2012, pp. 126-131.

[8] M. Jørgensen, "Top-down and bottom-up expert estimation of software development effort," *Information and Software Technology*, vol. 46, pp. 3-16, 2004.

[9] M. Shepperd and C. Schofield, "Estimating software project effort using analogies," *IEEE Transactions on software engineering*, vol. 23, pp. 736-743, 1997.

[10] B. Boehm, B. Clark, E. Horowitz, C. Westland, R. Madachy, and R. Selby, "Cost models for f uture software life cycle processes: COCOMO 2.0," *Annals of software engineering*, vol. 1, pp. 57-94, 1995.

[11] C. Abts, B. Clark, S. Devnani-Chulani, E. Horowitz, R. Madachy, D. Reifer, *et al.*, "COCOMO II model definition manual," *Center for Software Engineering, University of Southern California*, 1998.

[12] C. E. Walston and C. P. Felix, "A method of programming measurement and estimation," *IBM Systems Journal*, vol. 16, pp. 54-73, 1977.

[13] [Online]. Available: http://promise.site.uottawa.ca/SERepository. Accessed: Nov. 9, 2016.

[14] [Online]. Available: http://terapromise.csc.ncsu.edu.Accessed: Nov. 9, 2016.

[15] A. Sheta, D. Rine, and A. Ayeshs, "Development of software effort and schedule estimation models using soft computing techniques," in *2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence)*, 2008, pp. 1283-1289.

LOCATION SERVER BASED PROACTIVE ROUTING PROTOCOL

Muhammad Usman Butt
MS (Computer Science)
Isra University, Hyderabad , Pakistan
usmanbutt@outlook.com

Arshad Shaikh
Assistant Professor
Isra University, Hyderabad , Pakistan
arshad.shaikh@isra.edu.pk

Dr. Hameedullah Kazi
Pro Vice Chancellor
Isra University, Hyderabad , Pakistan
hkazi@isra.edu.pk

***Abstract*— Routing protocols in MANETS have become a research focused field with the rising trend of wireless communication. The two most important types of routing protocols in MANETs, that fall into the category of topology based routing protocols are Reactive and Proactive routing protocols. Reactive protocols have lower routing overhead since routes are constructed on demand. This type of routing creates routes only when required by the source node. In Reactive protocols, every node need less amount of information, as compared to Proactive Protocols. The route request message is flooded all over the network to find the shortest path to the destination. In case of high mobility MANETS, this type of flooding consumes a huge amount of network resources and encounters a high cost. The proactive protocols on the other hand, construct routes from each node to all other nodes whether required or not. This is achieved by periodic exchange of routing data between the neighboring nodes. This periodic exchange of data also encounters a cost on the network. Both types of protocols suffer from performance and quality issues due to high mobility of network topology. This problem leads to the need of a position based routing protocol that overcomes the flaws of traditional topology based routing protocols.**

This paper presents a new position based routing protocol that is proactive in nature. This new protocol is simple and provides reliable as well as efficient routing for certain applications of MANETS. The new protocol is named Location server Based Proactive routing (LSPR) protocol and it is compared with some of the most famously known existing routing protocols namely LSAR and DSDV. Simulation results suggest that LSPR outperforms DSDV and LSAR when tested under different stress conditions.

Keywords*— *MANETS , DSDV , GPS , Reactive and Proactive Protocols.

1. INTRODUCTION

A Mobile Ad-Hoc Network is a collection of mobile nodes that are connected through wireless links. It is a transitory network without having any fixed access point, infrastructure, or centralized administration. In other networks like GSM networks, there are very few Base Stations and hubs join to an outside framework like Internet or Intranet utilizing an entrance point and geographically isolated into cells. There is a base station for every cell. Each Mobile Station connects particularly to the Base Station. But in MANETs every node works as both router as well as a host. The system topology is changing on the basis of the fact that the system among the nodes may contrast with time in view of nodes joining and leaving the network.

The main focus of MANET coordinating tradition is to make sense of right and gainful course between a few nodes with the objective that on time and reliable messages delivery is possible. Discovery of routes should be obtained with slightest overhead. Every node contains the topology information of the network in Proactive routing protocols as guiding tables by periodically exchanging routing information. Routes are created in reactive routing protocols, only when needed by the source node but proactive routing protocols require every node to keep up routing tables to gather coordinating information, and they react to changes in network topology by applying updates all through the network to keep up a dependable view of the system.

Position based routing protocols work better for ad hoc networks generally for two reasons: there is no need to keep routing tables updated and no need to have a large view of the network topology and its changes that results in reduced routing overhead.

There are various challenges in MANETs such as routing, dynamic topology, scalability, bandwidth optimization. But the major challenge in MANET is link failure due to high mobility. Topology-Based routing protocols become unsuitable for MANET when the nodes are highly mobile and topology changes dynamically. Position based routing protocols are regarded as efficient and scalable when mobility is high. Therefore, position based routing protocols emerged as a new attraction in the field of routing protocols for MANETS.

In Position based routing protocol routing is done in a node to node mechanism to forward the data packets. They are designed for huge networks having hundreds or thousands of nodes. The plus point of position based routing protocols is that they only rely on local information to forward the data packet, rather than to keep the information of the whole network.

The rest of the paper is organized as follows: next section (Section II) presents a review of some of the important Topology Based and Position based routing protocols for MANETs. Section III describes the methodology. Section IV presents Results and Discussion while Section V presents the conclusion followed by Section VI that presents future work.

2. LITERATURE REVIEW

The traditional routing protocols mainly depend on routing table's information which are responsible to accumulate routes to various possible directions. In order to ensure that the routing tables will be up-to-date and their reflecting of the real network topology, the nodes often exchange route updates and re describe the paths. They are classified into two different categories:

i. Link State (LS) algorithms.
ii. Distance Vector (DV) algorithms.

Following are the characteristics of Traditional Protocols:

i. One of the most important built-in characteristic of MANETs is Dynamic topology but conventional protocols are static.
ii. Traditional conventions are very reliant on occasional messages to process the progressions in the system topology. Whereas the node thickness increases, it obliges continuous trade of data between different nodes.
iii. Traditional protocols will be composed in such a route, to the point that they accept that bi-directional connections exist.
iv. Traditional protocols experience the ill effects of routing circles which are to be dodged in MANETs as they would make an unnecessary activity and the data transmission would be worthless.

I. Link State Routing

In this type of routing protocols, each node keeps a root of the complete topology with an expense for each one connection [1]. Every node in the system occasionally shows the connection expenses of each one friendly connection to the various nodes utilizing flooding. This type of flooding of connection expenses permits the system hosts to have these expenses predictable. Every host in the system has a directing table that is utilized to save all connection costs which a node gets. In the wake of getting the overflowed message, every node up-dates its routing table and chooses the most limited way to every objective node. These connection expenses can give wrong connection cost data at any node because of spread postponement, parceled systems, and so forth, which brings about a development of directing circle. These loops are fleeting, on the grounds that they vanish when bundle time-to-live terminated. Then again, these loops make an overhead in the system. Moreover, keeping up these directing tables make these conventions entangled.

II. Distance Vector

In this type of routing, every node just screens the expense of its friendly connections, however as opposed to broadcast this data to all hosts, it occasionally shows to every one of its neighboring hosts an appraisal of the shortest distance to each of the nodes in the system [1]. The getting node then utilizes this data to recalculate the routing tables, by utilizing a shortest way calculation.

Contrasted with LS, DV is more effective, less demanding to execute and obliged a great deal less storage room. Then again, distance vector can result in the development of both brief and extensive steering loops. The essential driver for this is that the nodes pick their home trust in a totally dispersed way focused around data that can be rigid.

2.1 Topology based Routing Protocols

Topology-based routing protocols apply information about links that are already in the network to perform packet forwarding. Previously known routing protocol were based on the information about the topology that consisted of route establishment and route maintenance. They used link information that existed in the network for packet forwarding. Topology based protocols are further divided into:

2.1.1 Proactive Routing Protocols:

In these kind of routing protocols every host/node present in the network holds data in the type of tables and this table needs to be updated if any change occurs in the network topology [2]. In this sort of protocol, all nodes in a system

keep up one or more tables of routes which are frequently updated. If system topology changes every node broadcast an information message to the whole system. But, it acquires extra overhead cost because of keeping up to date data and therefore; throughput of the system may be influenced yet it gives the real data to the accessibility of the system.

2.1.1.1 Destination Sequenced Distance Vector (DSDV)

DSDV protocol is a proactive routing protocol. Routing tables are maintained by every node. That routing table is continuously updated. Nodes can send or receive data to other nodes in a network with the help of routing information. Sequence numbers are basically originated from receiver node itself which assures loop freeness. Install time is for deleting fake entries from table. Authentic data is a pointer to a table containing information on how authenticate a route is and also used to damp fluctuations in network [3].

2.1.1.2 Optimized Link State Routing (OLSR) Protocol

OLSR is an optimization over a pure hyperlink state protocol by compacting the scale of know-how dispatched within the messages, and through reducing the quantity of retransmissions to diffuse these messages in the complete community. For this rationale OLSR protocol makes use of multi point relaying technique to successfully and economically flood it's manipulate messages. It presents choicest routes in phrases of number of hops, which are instantly available when wanted. OLSR is best compatible for significant and dense mobile networks [4]. OLSR does now not generate extra manipulate traffic in response to hyperlink disasters and additions. It's as a consequence suitable for networks with a high price of topological changes. Extra dense and massive a network is, extra optimization is executed as compared to the traditional link state algorithm.

2.1.1.3 Fish eye State Routing (FSR) Protocol

FSR introduces the proposal of multi-degree fish eye scope to minimize routing update overhead in significant networks. Fish eye state routing protocol makes use of the fish eye technique proposed to scale down the size of understanding required to symbolize graphical data. The consideration of a fish catches with high component the pixels close to the point of convergence. The subtle element diminishes as the crevice from the point of convergence increments. Arranged on this motivation, cell hubs trade upgrade messages all the more likely with closer cell hubs, and not more by and large with more distant hubs. The precision of the hubs comprehension will rely on upon how a ways a hub is. The hub data is additional exact whether it is closer, whilst the hub information turns out to be less precise in the event that it is more distant [5].

2.1.2 Reactive Routing Protocols:

In this type of protocols, every node need less amount of information, as compared to table driven protocols. As routing information of all nodes is not required in a community and intermediate nodes must not make routing choices [6].

Reactive Protocol has minimize overhead considering routes are decided on demand. If a node is wanting a path to the receiver, it starts a path discovery process in that network. If a path is found or all possible path finding conditions have been inspected than this discovery process is said to be completed. In case if a route for data transfer is established, it remains active and maintained until data is completely sent.

2.1.2.1 Ad-hoc on demand distance Vector (AODV)

AODV utilizes a course revelation procedure to progressively fabricate new courses on an as need premise. AODV is a disseminated calculation utilizing separation vector calculations, for example, the Bellman Ford calculation. At the point when a course to a destination is obscure, AODV makes a course ask for bundle and communicates it to its neighbors. The fundamental preferred standpoint of this convention is that courses are built up on interest and destination succession numbers are utilized to locate the most recent course to destination. Along these lines association setup postponement is less. But since courses are just kept up while in use, it is ordinarily required to play out a course disclosure before parcels can be traded. This prompts a postponement for the main bundle to be transmitted.

2.1.2.2 DSR

Dynamic Source Routing (DSR) is a steering convention for remote lattice systems. It is like AODV in that it frames a course on-interest when a transmitting hub demands one. In any case, it utilizes source directing as opposed to depending on the steering table at every middle gadget.

2.1.3 Hybrid Routing Protocols:

The blend of table driven and on interest results in half and half conventions [7]. These conventions convey an improved exchange off between correspondence overhead and postponement, yet this exchange off is subjected to the measure of a zone and the elements of a zone. The on-interest strategy is utilized when data between various zones are steered. Along these lines, a way to a collector in the identical zone is built up immediately, while a way discovering process and a way upkeep technique is required for recipients which

are available in various zones. The thought used in this convention can convey an unrivaled tradeoff between correspondence overhead and deferral, however this exchange off is dependent on the zone and the elements of that zone. Moreover, half breed approaches give a trade off on adaptability problem in connection to the recurrence of end to end association, the aggregate number of hubs, and recurrence of change in topology. Along these lines, this methodology is a proper possibility for steering in a substantial system. Zone Routing Protocol (ZRP) is a famous type of Hybrid Routing Protocol.

2.2 Position Based Routing Protocols:

Position based steering convention uses position data to find the careful areas of destination hub and additionally its neighbor hub It utilizes position data to give more dependable and also proficient directing for specific applications and this data is for the most part gotten through Global Positioning System (GPS) and area administrations. By the utilization of area administrations and sending procedures its execution is vastly improved than topology based directing convention. It show better versatility, vigor against incessant topological changes. These steering conventions are intended to enhance proficiency and execution of the system. Steering is done in a jump by-bounce style to forward the information bundles. It is intended to handle organizes that have numerous hubs. One favorable position of this sort of steering is that it is completely in view of nearby data to forward the information parcel, as opposed to keep the system wide data. This will prompt highly diminished directing overhead and expansion the bundle conveyance rate. Position data of every hub is dictated by the utilization Location Services and Forwarding procedures are utilized to forward the information bundles [8]. At the point when a source hub needs to transmit a bundle to the destination, it ought to get the area (x, y) of the destination through a location service.

2.2.1. Location Aided Routing (LAR) Protocol

Host portability can bring about incessant unusual topology changes. Along these lines, the undertaking of finding and keeping up courses in MANET is nontrivial. Location Aided Routing (LAR), as its name proposes, makes utilization of location data to decrease steering overhead. Location data utilized as a part of the LAR convention might be given by the Global Positioning System (GPS). With the accessibility of GPS, it is workable for a portable host to know its physical area. LAR limits the quest for a course to the alleged solicitation zone, decided in view of the normal area of the destination hub at the season of course disclosure. Location Aided Routing (LAR) conventions restrict the quest for another course to a littler "solicitation zone" of the specially appointed system. . Demand zone is the zone in which the hub advances the course ask for just when the hub is inside the zone. At the point when the hubs does not has a place with solicitation zone then it basically disposes of the message. Expected zone is the territory in which there is the most extreme likelihood of finding the destination hubs. This outcomes in a noteworthy diminishment in the quantity of directing messages.

2.2.2. Distance Routing Effect Algorithm for Mobility (DREAM) Protocol

DREAM convention which additionally utilizes the hub area data from GPS frameworks for correspondence. DREAM is a section proactive and part receptive convention where the source hub sends the information parcel "in the heading" of the destination hub by particular flooding. The sender will communicate the parcel towards hubs in a constrained part of the system; to all single jump neighbors towards the destination. DREAM calculation is a proactive convention that uses a constrained flooding of area redesign messages Since DREAM utilizes the confined directional flooding to forward information bundles themselves, there will be various duplicates of every parcel in the meantime. This builds the likelihood of utilizing the ideal way; be that as it may, it diminishes its adaptability to expansive systems with a high volume of information transmissions and makes it more reasonable for applications that require a high unwavering quality and quick message conveyance for rare information transmissions.

2.2.3. Location aided Knowledge Extraction Routing (LAKER) Protocol

Location aided Knowledge Extraction Routing (LAKER) convention gains from past activities to guide future practices. LAKER can step by step find current topological qualities of the system, for example, populace thickness dissemination and movement load status. This learning can be sorted out as an arrangement of managing courses, each of which comprises of a chain of directing positions between a couple of source and destination areas. The directing course data is found out by individual hubs amid course revelation stage, and it can be utilized to guide future course disclosure forms in a more proficient way. LAKER is particularly appropriate for versatility models where hubs are not consistently dispersed. LAKER can misuse topological attributes in these models and

cutoff the hunt space in course revelation forms in a more refined granularity than LAR convention [9]. The LAKER convention minimizes the system overhead amid the course disclosure process by diminishing the zonal range in which course ask for bundles are sent. Amid this procedure, LAKER separates learning of the nodal thickness appropriation of the system and recall a progression of ""essential"" areas on the way to the destination. These areas are named "controlling courses" and with the assistance of these directing courses the course disclosure procedure is limited down.

2.2.4. Location Server Assisted Routing Protocol

LSAR (Location Server Assisted Routing) Protocol is a reactive protocol which utilizes Geographical data to locate the briefest way between the hubs. Rather than direct flooding, this convention send information bundles through connections subsequently having the highlight of topology based convention. LSAR convention is reactive in nature that is root gesture will make the course just when required. In LSAR root node is in charge of route support. It send Root Announcement message to all hubs intermittently and all hubs upgrade their directions once they get the message. This all courses are overhauled at root hub. LSAR has a very long set of actions to perform data forwarding that are : Send Rout request, Receive Route Request, Send Route Reply, Receive route reply, Send Route set, Receive route set which culminate in heavy delay in route discovery process. LSAR was compared with AODV and LAR and LSAR performed better than both of them.

3. METHODOLOGY

This section includes implementation of LSPR as well as its comparison with its competing counter parts. Location server based proactive routing (LSPR) protocol is a hybrid protocol applying the approach of being position based as well as proactive protocol (because it uses links to send and receive data packets). Because it is proactive in nature it means each and every node has the routing data whether needed or not.

3.1 Working of LSPR:

This protocol is implemented intending to curtail the number of routing messages exchanged by introducing a special node in the network, called Location Registrar. Initially node 0 in the network is selected as location registrar or root node. The location registrar maintains the routing information for all the nodes in network. Unlike AODV in which Shortest path is established based on hop counts, while in our algorithm shortest path is established based on GPS location of nodes.

In LSPR each node updates the location registrar about its GPS coordinates, in reply to the root announce message, and registrar sends adjacency matrix to every node about their neighboring nodes while sending root announce message. When a route is required to any node, the node applies the shortest path algorithm on the available data in adjacency matrix to construct the shortest path and pass the data to the destination node.

Ideally, only one location registrar is present in the entire network but this is only true if the network is quite stable so that all nodes get enough time to get to know about the root's presence. When a route is required, all nodes know route of every other node. Thus will send data packet directly to the destination.

No flooding is done in LSPR. Unlike LSAR, in LSPR root node is not responsible for finding path for every node by using shortest path algorithm. But every node finds its own shortest path. Thus reducing routing overhead for root node as well as the whole network.

A major problem that rises in position based routing protocols is that they cannot differentiate between two alternatives:

One via a node that is about to go beyond range, leaving a broken link and one via a node that is closer and carries a lesser risk of broken link. LSPR overcomes this problem quite efficiently by reducing the node range described for distance formula from the antenna range. (i.e. node range in distance formula is 100 meters and antenna range is 120 meters). By doing this LSPR gets a margin in range, if a node is between 90 and 100 meters away from another node, it will still be considered in range but if it move ahead of 100 meters range it will still be in antenna range and will transmit data and will not culminate in loss of data and will get almost 20 meters margin which is quite tremendous.

Unlike LSAR, LSPR only has 3 steps to forward data to the destination.

Following are some main steps in LSPR:

Step 1: Root Announce

As LSPR works with Root Node or location registrar. For convenience we have selected node 0 as root node. First step in our protocol is Root Announce. Location registrar or root node will send a Root Announce Packet to all nodes present in the network when simulation start, this root announce packet also contains adjacency matrix containing adjacent nodes to every node in the network. Initially the adjacency matrix in first root announce packet is empty, but it is filled as other nodes send their GPS coordinates to root node via announce to root packet. The root announce packet is received and forwarded by all nodes in the network as shown in Figure 1.

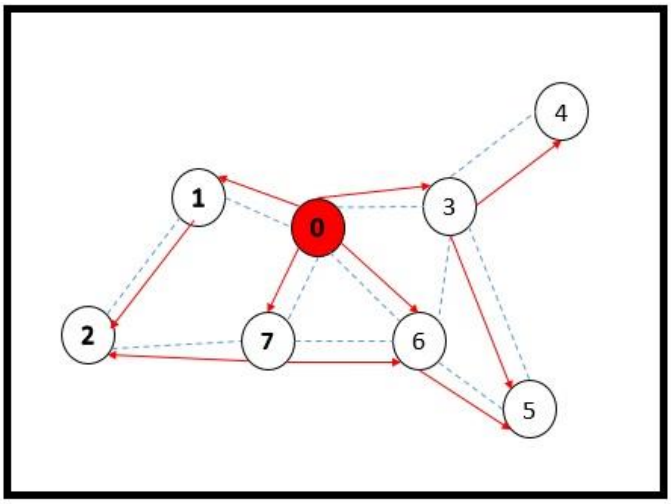


Figure 1. Root Announce

Step 2: Announce To Root

When all nodes received root announce message, the next step is Announcing to Root. Basically all nodes will send their locations to root node in reply of the root announce message (as shown in Figure 2) and root will register them by creating an adjacency matrix. The root node performs distance formula on the GPS coordinates and fills adjacency matrix with 1's and 0's.That is how location of all nodes is saved. Through this all nodes will also know that for root node which node is their next hop and this is how all nodes are informed about their adjacent nodes.

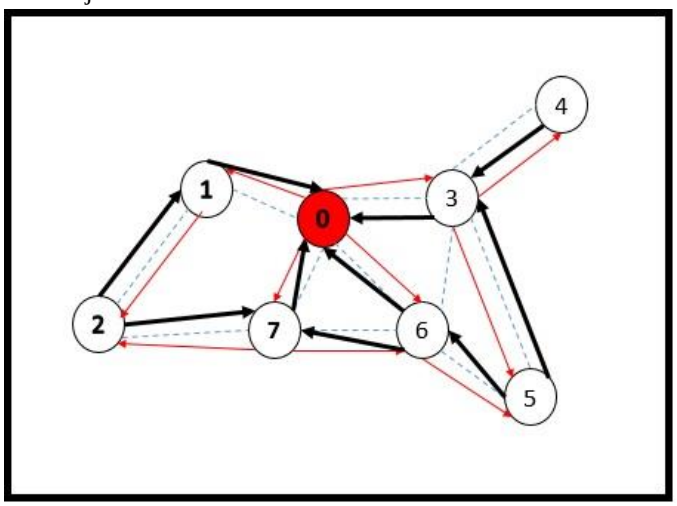


Figure 2. Announce To Root

Step 3: Data Forwarding

When every Node Receives adjacency matrix it means it can now forward data to any node in the network. For example if a route is required from node 7 to node 4 (as shown in Figure 3), node 7 performs Dijkstra algorithm on the given adjacency matrix and gets the complete and shortest route to the destination node that is 7 to 6 to 3 to 4 and thus forwards the data.

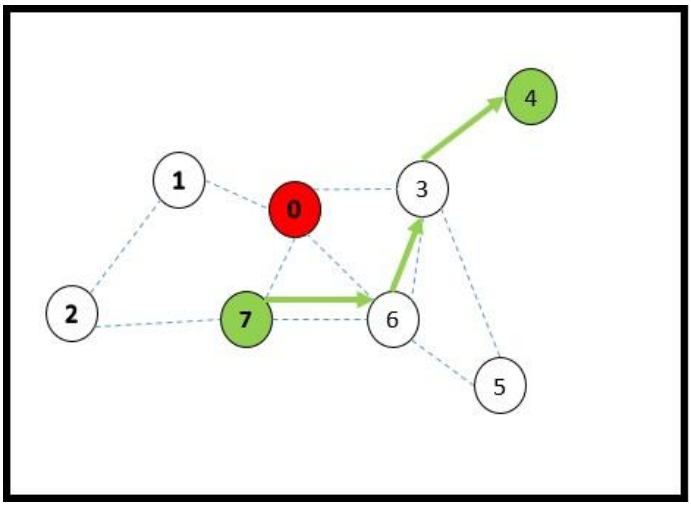


Figure 3. Data forwarding

3.2. Performance Evaluation

We have implemented LSPR using C++ and simulated in NS-2, one of the most general open source simulator to analyze the performance for LSPR as well as its competing counter parts DSDV and LSAR. Three different Quality of Service Parameters were used to evaluate the overall performance of LSPR, LSAR and DSDV.

3.2.1 QOS Parameters

a)Packet Delivery Ratio

Packet delivery ratio is ascertained by partitioning the quantity of packets received by the quantity of packets sent. CBR packets are utilized for genuine correspondence. PDR of a routing protocol ought to be as high as it could be for best execution.

b)Throughput

Throughput is the amount of data packets moved successfully from one node to another in a unit time, typically measured in bits per second (bps). Throughput and PDR actually go hand in hand. So if PDR increases, Throughput also increases and vice versa.

c)Average End to End Delay

The time taken for packets to be transported over a network from sender to receiver is called End to End Delay. For the better execution of the protocol Average end to end delay must be as less as could be expected under the circumstances.

It is characterized as the time a data packet is gotten by destination minus the time the packet is produced by the source.

3.2.2 Simulation Parameters

In MANETS, Different parameters are required to comprehensively simulate routing protocols. These parameters are used to describe the characteristics and behavior of ad hoc network. Table 1 shows the simulation parameters used in the simulation:

PARAMETERS	VALUES
TRANSMISSION RANGE	120 M
SIMULATION TIME	600S
TOPOLOGY SIZE	500M X 500M
NUMBER OF MOBILE NODES	50
MOBILITY MODEL	RANDOM WAYPOINT
TRAFFIC TYPE	CBR (512 BYTES PACKET)
TRAFFIC	5, 10, 20 (SIMULTANOUS CALLS)
MOBILITY SPEED	5, 10, 15, 20, 25 M/S
NS-2 VERSION	NS-2.35

Table. 1 Simulation Parameters

4. RESULTS AND DISCUSSION

In this section LSPR is compared with LSAR and DSDV to determine the protocol that performs better under different stress conditions for the following QoS metrics:

A. Packet Delivery Ratio

As discussed earlier packet delivery ratio describes the ratio of amount of packets successfully received at the destination. As a matter of fact a protocol is said to be better if its PDR is better than others. Figure 4 shows the packet delivery ratio of LSPR, DSDV and LSAR under all 3 stress conditions 5,10 and 20 simultaneous calls for 5 mobilities (5,10,15,20,25 m/s). For 5 simultaneous calls all three performed better but as call load increased to 10 and 20 simultaneous calls, the PDR of LSAR and DSDV dropped significantly at 50 and 60 %. While PDR of LSPR remained at averagely 80 to 90 %. So in terms of packet delivery ratio LSPR remains ahead of both LSAR and DSDV.

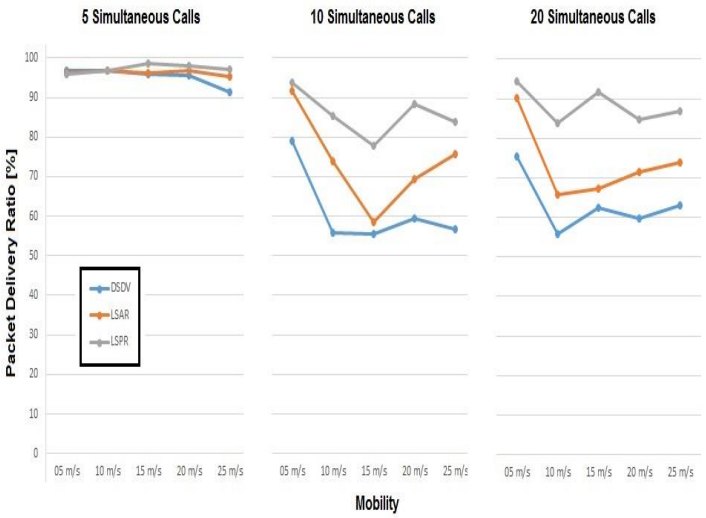


Figure 4. Packet Delivery Ratio

B. Throughput

Throughput is the amount of data packets received per unit amount of time. Throughput of LSPR, DSDV and LSAR can be seen in Figure 5. PDR and throughput are directly proportional to each other that means if one increases, the other increases too and vice versa. So graph of both PDR and throughput look similar. At 5 simultaneous calls throughput of all three protocols was at 40 kbps but as the call load increased to 10 and 20 simultaneous calls, LSPR outperformed both LSAR and DSDV and yet proving to be better than both of them in terms of throughput as well.

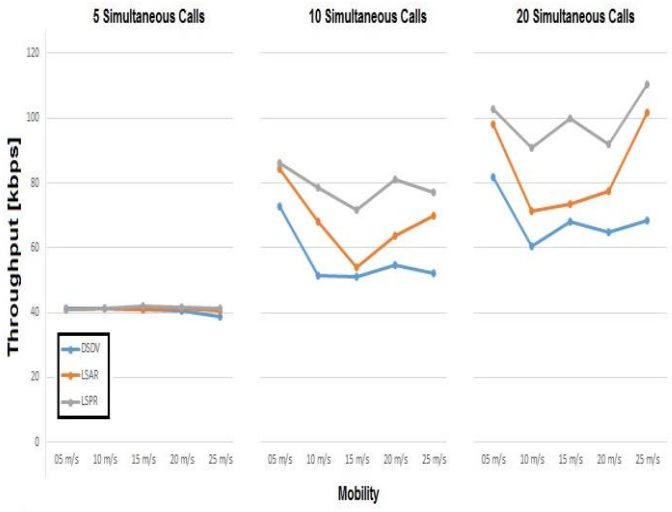


Figure 5. Throughput

C. Average End To End Delay

Average end-to-end delay for LSPR is almost same if we see the average values for every stress condition as shown in Figure 6. However for DSDV and LSAR the average end-to-end delay is increasing as the number of simultaneous connections are increasing and they turned out to be performing poorly in terms of average end to end delay.

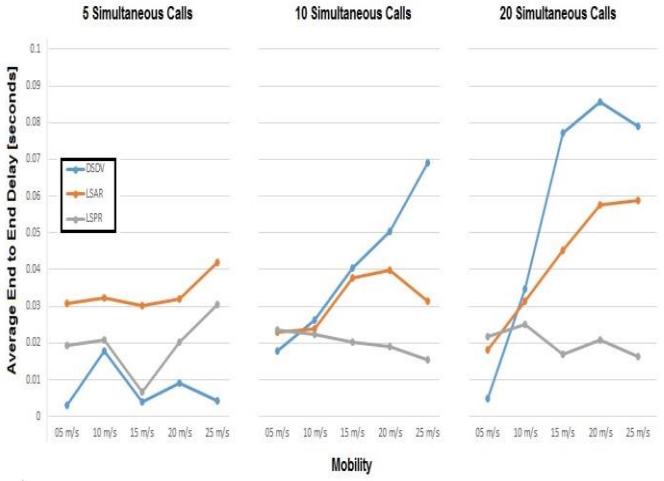


Figure 6. Average End To End Delay

In all of these QoS metrics one thing is certain about LSPR and that is the mobility is not effecting much on LSPR, as a matter of fact LSPR is performing even better in most of the high mobility conditions and other stress conditions we have applied, which is a good sign and we can say that LSPR outperforms DSDV and LSAR in almost all aspects.

5. CONCLUSION

A novel position based routing protocol LSPR is developed and implemented for Mobile Ad-hoc Networks. The protocol is simple and generic and overcomes the shortcomings of existing position based as well as topology based routing protocols. As LSPR is a position based as well as proactive protocol, it is compared with a proactive protocol i.e DSDV and a position based routing protocol i.e LSAR. Simulation results suggest that LSPR outperforms both DSDV and LSAR under different stress conditions and different mobility speeds in terms of all three quality of service parameters PDR, Throughput and Average End to End Delay.

6. FUTURE WORK

As future work we suggest, introducing a root node election mechanism in LSPR that will immediately elect a new root node in case of a dead root node. We have used Dijkstra algorithm as a shortest path algorithm in LSPR, another suggestion would be to implement LSPR using any other latest shortest path algorithm to further enhance the performance of LSPR.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to Mr. Arshad Shaikh for his continuous support of my MS (CS) research and for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research. Besides my supervisor Dr. Hameedullah kazi, I would like to thank him for his encouragement and insightful comments. I would like to thank my parents for supporting me in every possible way throughout my life.

REFERENCES

[1] Santiváñez, César A., Ram Ramanathan, and Ioannis Stavrakakis. "Making link-state routing scale for ad hoc networks." Proceedings of the 2nd ACM international symposium on Mobile ad hoc networking & computing. ACM, 2001.

[2] Hemagowri, J., C. Baranikumari, and B. Brindha. "A Study on Proactive Routing Protocol in Ad-hoc network." International Journal of Modern Engineering Research (IJMER). pp 1 4 (2013).

[3] He, Guoyou. "Destination-sequenced distance vector (DSDV) protocol." Networking Laboratory, Helsinki University of Technology (2002): 1-9.

[4] Clausen, Thomas, and Philippe Jacquet. Optimized link state routing protocol (OLSR). No. RFC 3626. 2003.

[5] Chiu, Ting-Hung, and Shyh-In Hwang. "Efficient fisheye state routing protocol using virtual grid in high-density ad-hoc networks." 2006 8th International Conference Advanced Communication Technology. Vol. 3. IEEE, 2006.

[6] Pandey, Ashwini K., and Hiroshi Fujinoki. "Study of MANET routing protocols by GloMoSim simulator." International Journal of Network Management 15.6 (2005): 393-410.

[7] ShrutiVaish, Amit Kumar Gupta, Shilpi Singh, and Saroj Sharma. "Hybrid Routing Protocol."

[8] Mauve, Martin, Jorg Widmer, and Hannes Hartenstein. "A survey on position-based routing in mobile ad hoc networks." IEEE network 15.6 (2001): 30-39.

[9] Li, Jian, and Prasant Mohapatra. "LAKER: Learning from past actions to guide future behaviors in ad hoc routing." Wireless Communications and Mobile Computing 7.4 (2007): 495-511.

Assistance of K-Means in Non-Intrusive Load Monitoring

Farah Andleeb Siddiqui
Electrical Engineering Department
DHA Suffa University
Karachi, Pakistan
siddiquifarah19@gmail.com

Syed Daniyal Fatmi
Transmission Department
K-Electric
Karachi, Pakistan
daniyalfatmi@hotmail.com

***Abstract*– The energy crisis throughout the world has imposed a solemn duty on the energy users as well as on the utilities providing the energy to reassess their energy demand utilizing the load monitoring techniques and to introduce new means of energy efficiency and conservation throughout the masses. Energy management cannot be achieved through individual effort but requires the collective effort of both the energy provider and consumer, electricity audits have therefore gained an immense popularity whether it be a residential, commercial or industrial sector.**

Load monitoring is the building block of energy management plan. The advancement in communication have made it fairly easy to access the information even from remote areas through Bluetooth, satellite communication and internet. However, the information access is not possible without the installation of large number of intrusive and expensive sensors. The efficiency, reliability and accuracy is adversely affected by the employment of many sensors compared to relatively few sensors.

Non-intrusive load monitoring technique provide an ideal platform for the extraction of useful information with minimum number of sensors that increases its reliability and lowers the cost of installation. Its most important feature is the use of various artificial intelligence techniques that can be modified over the time and according to needs and usage.

This paper discusses the Non-Intrusive Load Monitoring concept in general laying the emphasis on the disaggregation algorithm technique used for the project and highlight its result.

Keywords—non-intrusive load monit*oring; signature space; classification algorithm; k-means clustering.***

I. BACKGROUND TO NON-INTRUSIVE LOAD MONITORING

The concept of NILM was pioneered by George W. Hart in 1982, a professor at Massachusetts Institute of Technology. Non-intrusive load monitoring is designed to monitor that electrical system which has a number of individual devices that switch on and off independently [1-3].

In this way an aggregated waveform of the energy information is obtained via single sensor attached at the main electrical board. Through various refined analysis of total load NILM estimates the nature and number of individual loads along with their consumptions. Since only a single sensing point is available so NILM provides a convenient and effective method in contrast to the traditional method. There are two major reason for the name 'non-intrusive':

1. Firstly, for the load detection no physical or electrical intrusion is created in the system.
2. Secondly no intrusion in the personal territory during the data collection phase.

Recently this method is gaining popularity due to the advancement in embedded systems as well as in field of digital computation. In general, NILM could be defined by two of its major components; **appliance signature** and **classification algorithm** [4].

The concept of NILM could be summarized in the following equation:

$$P_i(t) = \sum_{k=1}^{n} P_k(t) \quad (1)$$

Where P_i is the total power consumed and is equal to the aggregate of 'n' individual device power P_k. the task of NILM is to decompose this power into its constituent devices.

NILM has an advantage over the traditional method that it has a simplified hardware and complex but efficient software to separate out the individual load from the aggregated waveform [5].

Figure 1 aids the NILM concept which depicts total real power versus time for the duration of around 20 minutes. During this time interval the step changes of different devices turning on/off is evident.

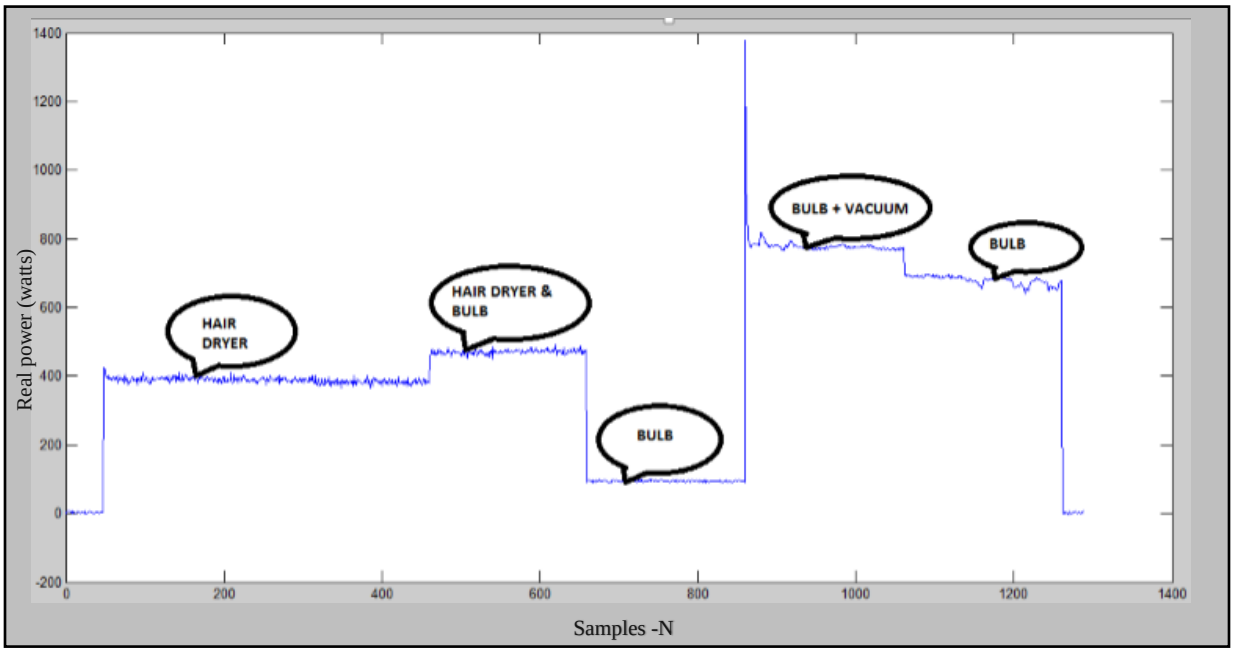


Fig. 1. Aggregated load waveform (obtained from original hardware)

II. PROJECT OVERVIEW

A. *Project Hardware:*

The hardware design consists of two portions:

- Data Collection Module: Data acquisition module is one of the most important parts of this project. To apply NILM algorithms a huge set of data is needed, therefore an accurate measurement circuit has to be made. Our data collection is done through an Arduino Energy Monitor and uses Emon library which measures main voltage and current. The key components of this energy monitor are CT sensor and voltage transformer, an interfacing circuit which interface CT[1] and VT[2] with Arduino due and SD card interfacing circuit.
- Data storage Module: The data acquired from different loads plays a vital role in the implementation of this project and throughout the project a series of extensive data from numerous loads has been extracted, thus we needed a storage device to handle large amount of data. To interface SD card with Arduino due we need a SD card breakout.

Figure 2 shows the hardware that was designed for the purpose of load monitoring.

B. *Load Selection:*

The loads selected were purely single state (on/off) loads and single phase. The project designed was suitable for residential load monitoring only. However, with the addition of further appliance features like harmonics, transients etc. the project could be alleviated for the monitoring of commercial and industrial building.

The load selected were bulb (100W), hair dryer (340W), iron (950W) and vacuum cleaner (700W).

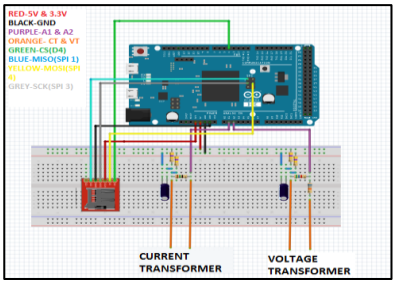


Fig. 2. Breadboard View of hardware on Fritzing

C. *Appliance Signature:*

The appliance signature is the foundation of NILM because it is these signature on which the disaggregation algorithm is built upon. Signature of an appliance differentiate it from many other devices. There could be a number of signature of a

device but for our project we selected the steady-state signature space which comprises of the real and reactive power of the appliance. The appliance signature does not overlap. Figure 3 shows the signature space having signature of all selected loads.

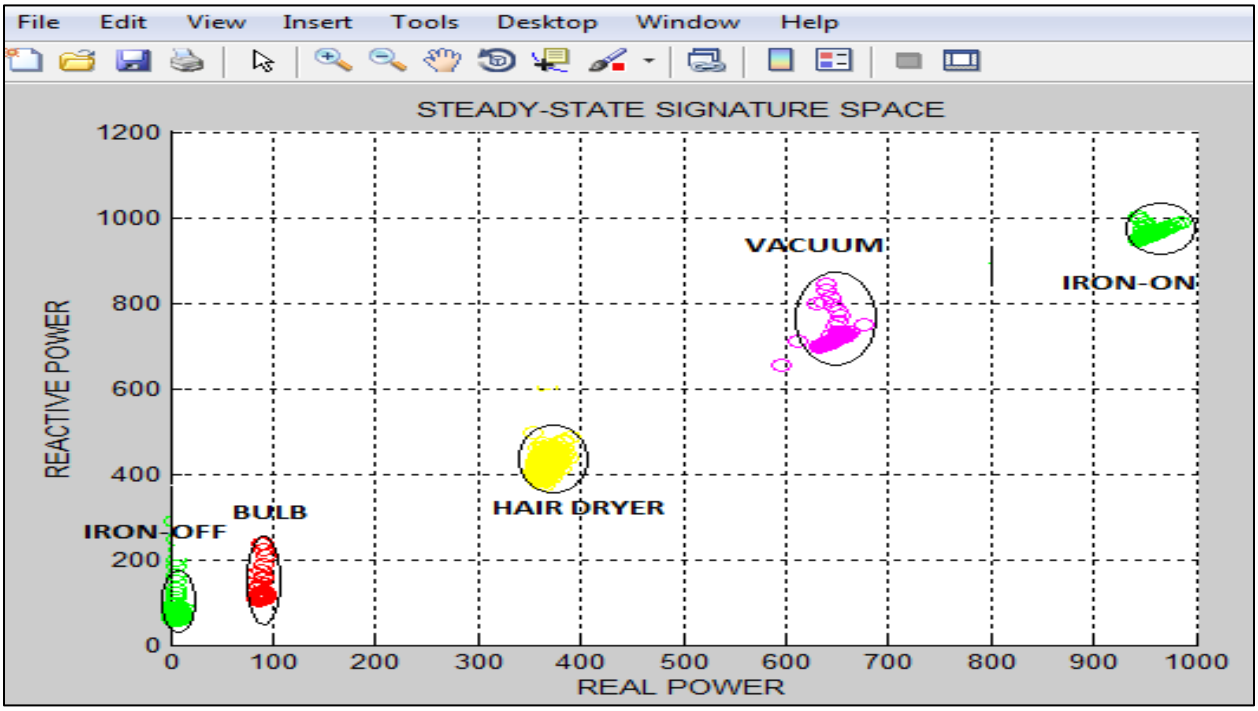


Fig. 3. Steady state signature space (obtained from original data)

III. DISAGGREGATION USING K-MEANS CLUSTERING

Once the data has been collected and transferred to the computer, the next step is to disaggregate the compiled waveform into individual loads using any supervised or unsupervised technique of artificial intelligence.

For the purpose of our project we used the K-means toolbox of MATLAB and adopted the supervised learning process. In supervised learning there are two steps of training and testing.

K-means is a clustering method in which the data is grouped into k group where k is the number of pre-chosen groups. The grouping is done by minimizing the sum of squared distances between the collected data and the centroid. (Euclidean distance). Classically, the centroid means the center of mass of a geometric object having uniform density. But in this algorithm centroid is the mean vector [6].

K-means clustering technique have been previously used by many researchers following different algorithm. K-means under some cases could be supervised but mostly it comes under the category of unsupervised learning [7].

In MATLAB we have an inbuilt function of K-means that could be used to find the centroid of any cluster and saved for the use of further processing. The flowchart of figure 4 explains what is happening at the back end of this function. K-means clustering is one of the easiest clustering technique that simplifies large input data into smaller clusters for further processing. For the purpose of our project; post processing of the data was adopted.

A. Weakness of K-Means:

There are certain drawbacks of k-means that are necessary to consider when using this technique. They are as follows:

- With fewer samples of data, initial grouping will determine the cluster significantly.
- The number of clusters, k, must be determined beforehand.

- With fewer samples of data, inaccurate clustering can occur.
- We never know which variable contributes more to the clustering process since we assume that each has the same weight.
- The accuracy of mathematical averaging weakens because of outliers, which may pull the centroid away from its true position.
- The results are clusters with circular or spherical shapes because of the use of distance.

The possible solution to these problems could be:

- Include as many samples of data as possible (the more data, the more accurate the results).
- To avoid distortions caused by excessive outliers, the median can be used instead of the mode [8].

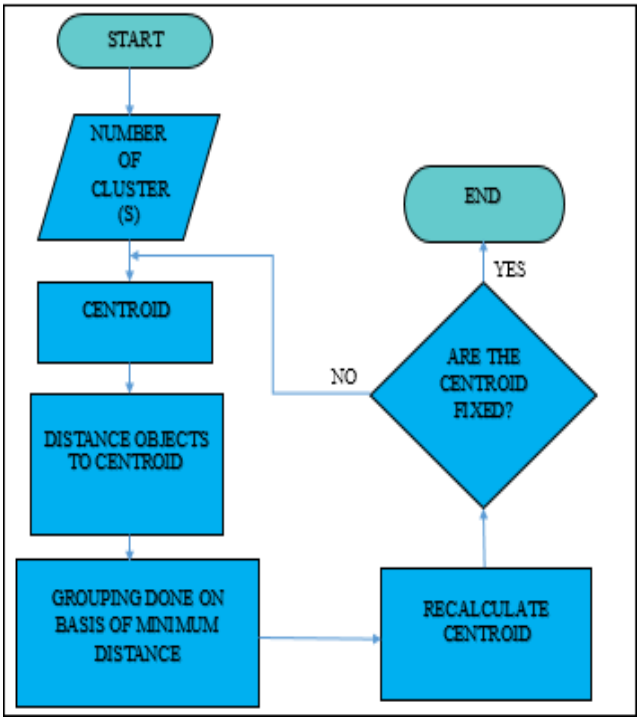
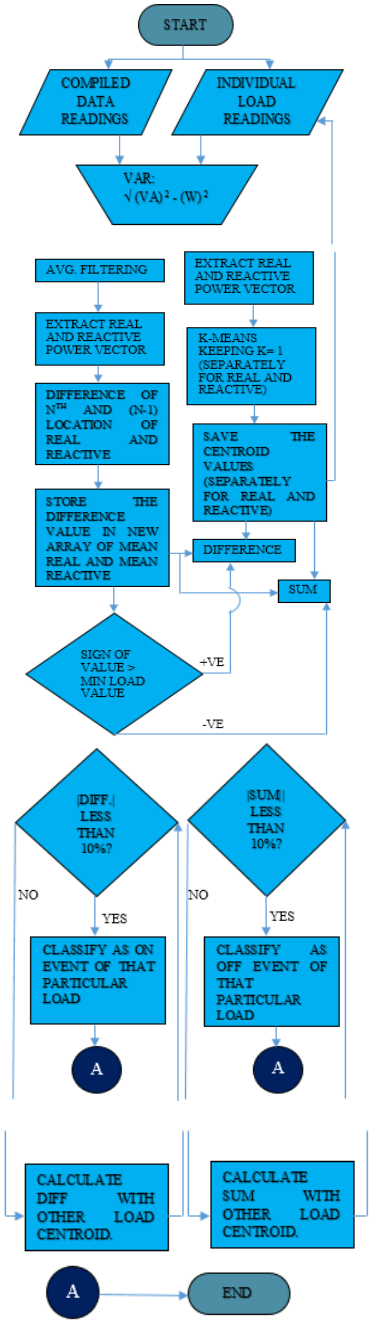


Fig. 4. General K-means Algorithm

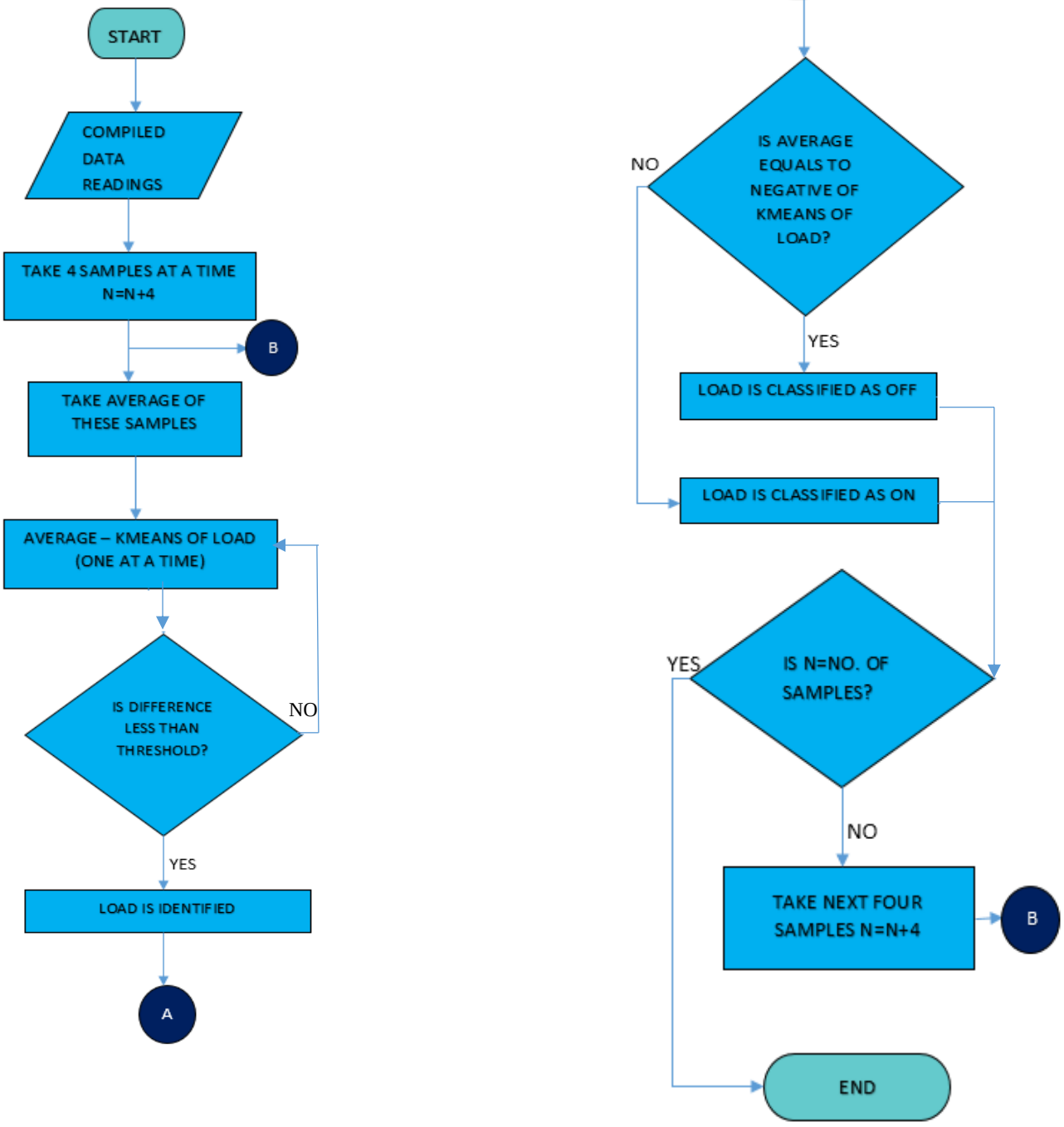
IV. DISAGGREGATION ALGORITHMS

There are two approaches that have been followed for the disaggregation of waveform into its ingredient devices. Both the algorithm designed are based on k-means clustering and are tested on the project prototype and the readings are the real data of the devices collected through our hardware module. The starting of the steps is almost same in both the algorithm but after that different paths have been followed. The following is the flow chart of the code that has been designed using k means.

A. First algorithm:



B. Second algorithm:



Both the algorithm discussed above have been tested and worked quite satisfactorily on the aggregated load waveform.

The training and testing phase of k-means algorithm is discussed in the later section.

V. TRAINING, TESTING AND RESULTS

During the training phase each of the load selected was connected individually on separate time intervals and the readings were recorded and saved. Important consideration was given to the fact that the device already reached its steady state before the data was being collected. The data was then transferred to MATLAB using SD card. The k-means centroid for each load was calculated and saved for later use. Since only one device was connected at a time during training so the number of cluster was just one.

Figure 5 shows clearly the cluster of individual devices with their k-means centroids marked.

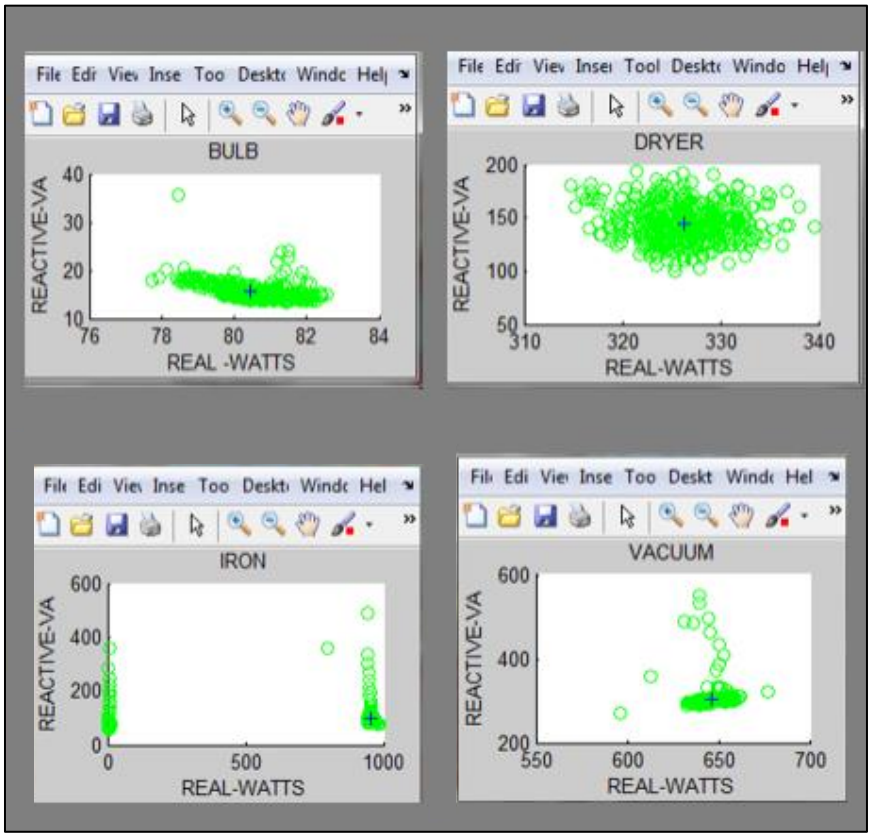


Fig. 5. K-Means Centroid

During the testing phase all the devices were connected at the same time and switched on and off in turns so as to get the aggregated load waveform as shown in figure 6. The complete data set was saved in the SD card and transferred to MATLAB for further processing. The data was passed through the above mentioned algorithm and the turning on and off time of a device with its name was effectively and efficiently decoded by the algorithm.

Thus by the use of artificial intelligence technique it was possible to achieve first level of automation in the field of load monitoring.

The blue waveform in figure 6 is the aggregated waveform and the overlapping green lines point out the step changes. The positive line indicates the turn on event and negative indicates vice versa.

The data tips are mentioned on every step change in the waveform and the same results are indicated on command window of the MATLAB as the algorithm runs on the aggregated data and verifies that the results are correct.

Table 1 shows the snapshots of MATLAB command window.

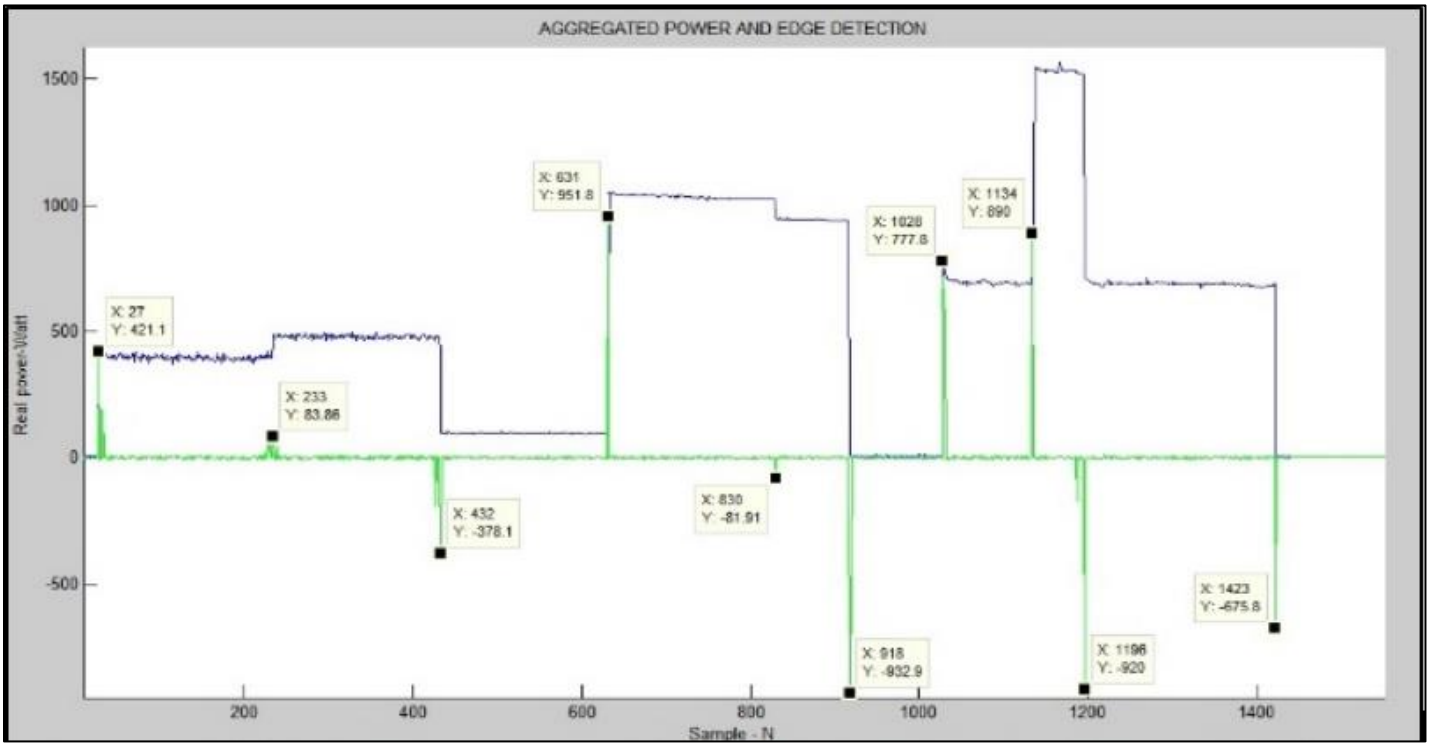
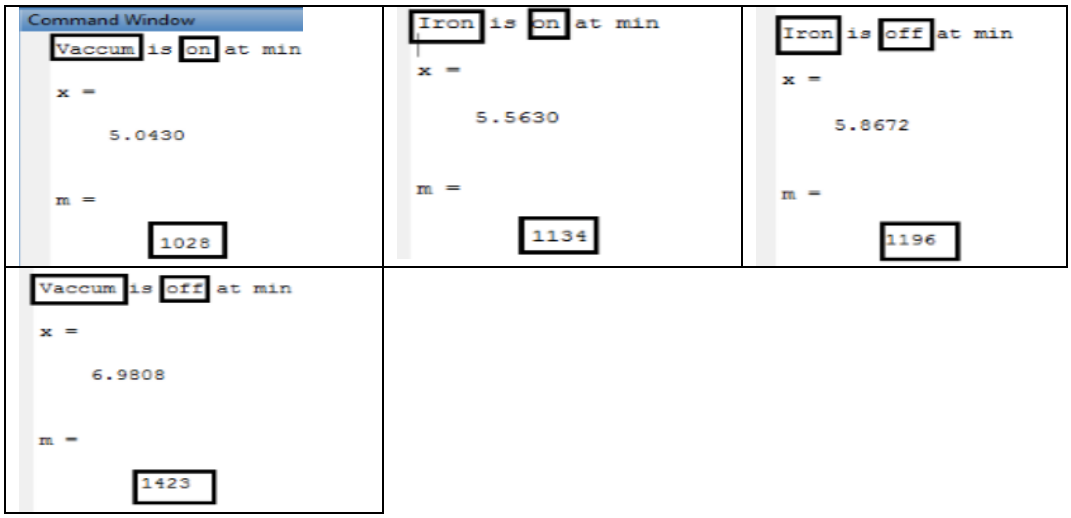


Fig. 6. Aggregated load waveform with edge detection

Table 1: MATLAB results

Command Window DRYER is on at min x = 0.1325 m = 27	Bulb is on at min x = 1.1430 m = 233	Command Window DRYER is off at min x = 2.1192 m = 432
Iron is on at min x = 3.0955 m = 631	Command Window Bulb is off at min x = 4.0717 m = 830	Iron is off at min x = 4.5034 m = 918



VI. CONCLUSION

NILM technique enchanted the domain of load monitoring for energy management leading to a more cost effective, reliable and efficient power system network. Since it suffices both the user and the supplier with the knowledge of electricity consumption patterns. This particularly benefits the sector of demand side management.

The project was an integration of hardware and software. The employment of non-intrusive load monitoring initiated with the formulation of prototype comprising of data collection module and interfacing circuitry. The parameters computed includes current, voltage, active power, reactive power and power factor. This was followed by the selection of required features from the sampled data and then the application of NILM on them. The application of this tool however, requires assistance of artificial intelligence since it cannot disaggregate the load pattern into individual devices without it. Thus by using K-means the aggregate load data was separated out into individual devices being turned on or off.

The project is just a globule in the field of energy audit and with improvised algorithms holds great potential for the future engineers.

VII. FUTURE WORK

Significant future work includes the development and testing of modified algorithms to explore various load signatures. Optimization can be done for the detection of constantly operating loads like clocks and multistate appliances such as washing machine. Moreover, the information from data collection module can be transferred to the computer via Wi-Fi. Synchronization can be done with industry standards. The device can be implemented for three phase circuits providing feasibility to industries. This is obtainable by replicating the hardware twice along with few modifications in the coding.

Currently the device is performing post processing. However, the incorporation of real-time software application would benefit in real time processing. The data will directly transfer to MATLAB and will be processed at the same time resulting in better utilization of resources.

NILM holds smart grid compatibility with the same motive of two-way communication between consumer and the utility, thereby it would be easier to monitor and improve energy consumption of loads.

REFERENCES

[1] L. Carmichael, “Nonintrusive appliance load monitoring system,” EPRI J., pp. 45-47, Sept. 1990.

[2] C. J. Andrews, “Nonintrusive monitoring device for electric loads in commercial buildings,” MIT Laboratory for Electrical and Electromagnetic Systems Technical Report, 1988.

[3] G. W. Hart, "Advances in nonintrusive appliance load monitoring," Proc. Inform. Automat. ***Conj,*** 1991.
[4] S. M. GEORGE W. HART, IEEE, "Nonintrusive Appliance Load Monitoring," *PROCEEDINGS OF THE IEEE*, vol. 80, p. 1870, 1992.
[5] Langhman, C.; Lee, K.; Cox, R.; Show, S.; Leeb, S.B.; Norford, L.; Armstrong, P. Power signature analysis. *IEEE Power & Energy Mag.* **2003**, *1* (2), 56–63.
[6] T.Kanungo; D. M. Mount; N. S. Netanyahu; C. D. Piatko; R. Silverman; A. Y. Wu., *"An Efficient k-means Clustering Algorithm: Analysis and Implementation."*
[7] T.Finley; T.Joachims; "*Supervised K-means Clustering."*
[8] J.P.Ortega; Ma. D.R.B. Rojas; Maria J. Somodevilla Garcia; *"Research issues on K-means Algorithm: An Experimental Trial Using Matlab."*

CT-CURRENT TRANSFORMER
VT- VOLTAGE TRANSFORMER
VAR-REACTIVE POWER
VA- APPARENT POWER
W- REAL POWER

Potentiometer Based Hand Gesture Control Of Robot Wirelessly

A Cost Effeective Approach for Developing Robot

Rohana Kamran
Biomedical Engineering Department, Hamdard Institute of Engineering & Technology, Hamdard University
Karachi, Pakistan
rohanakamran@gmail.com

Hafiz Muhammad Abdul Basit Ayubi
Biomedical Engineering Department, Hamdard Institute of Engineering & Technology, Hamdard University
Karachi, Pakistan
m.basit@hamdard.edu.pk

***Abstract*—This paper provides a deep insight of development of wireless controlling of robot. Usually, robots are made to perform different tasks that are usually unable to perform by human beings. Large numbers of robots are now being made for special purposes like for security reasons, some robots are made for improvisation the life of physically disabled persons (PDPs). These robots are task specific because they are programmed accordingly. This paper describes the wireless controlling of robot by hand gesture, as the hand moves in either direction; the robot will follow the same in a controlled manner. This robot is controlled wirelessly via RF Modules (RF transmitter and RF Receiver). The main advantage of this approach is the cost effectiveness. This approach is 85% (cost of sensor) cheaper in cost than the accelerometer based hand gesture controlling of robot. By using this approach, we achieved the same results that were desired and to control speed of robot we used a very simple and famous technique known as PULSE WIDTH MODULATION (PWM). The presented idea of the undersigned project is taken from literature but it is not only a cost effective approach but the speed of the robot can be controlled in any direction. The results have been achieved using a commercially available software tool.**

Keywords—Flexion, Extension, Abduction, Adduction, Potentiometer, RF Module (Tx and Rx) , Decoder IC, Encoder IC, PDPs, Arduino UNO, Motor Driver IC, DC Motors, PWM

I. INTRODUCTION

In the modern era of technology, many kinds of robots are being developed by scientists and researchers. Some of them are fully autonomous while some are semi-autonomous. These robots do not only play an important role in military departments or rescue departments (for firefighters) but they also play very important role in Health and Medicine department. The proposed project is also designed for paralyzed patients whose whole body is paralyzed except hands. Robots based on two different technologies are available in market one of them is wireless and other is wired but, due to advancement in gesture controlled devices people are attracted towards them. [1] Ccontrolling of robot through the hand movements will enhance the quality of independency of the paralyzed patients. The purpose of developing this robot is to communicate with the robot from a distance in a friendly environment. This robot follows the instruction made by PDPs hand as if PDPs flex hand, the robot will move backward, and when hand extends, robot will move in forward direction, and when hand tilts towards right, the robot follow the same direction and similarly when hand tilts left side, the robot moves in that direction. This robot follows four directions as in +y-direction, it will move forward, in –y-direction, it will move in backward direction, similarly in +x- direction, it turns right and in –x-direction, it turns left. The robot presented in this paper, is a wireless robot and it is made by using two Potentiometers (one for +x and –x directions and second one +y and –y directions) as a voltage level regulator, comparators to compare two different voltage levels, Encoder IC (M145026) to encode the output signal information and decoder IC (M145027) to decode that information, RF module that consists of Transmitter (Tx) and Receiver (Rx) for wireless communication and microcontroller Arduino UNO for controlling the robot.

II. RELATED WORKS

The paper [1] describes the development of wireless robot by using Accelerometer sensor and PMDC (Permanent Magnet DC Motors). Accelerometer sensors are very expensive and there are many other options that can be implemented and replaced accelerometer cost effectively and we can also control the speed of motors in any direction by PWM technique or by any other technique. The paper [7][8] focuses on the development of the robotic Arm by using Flex Sensor, ZigBee and 3 Servo motor connected to the Arduino Uno which is controlled by processing software and a computer mouse. These robotic Arms are cheap and easily available which makes it free from unnecessary wire connection, reducing its complexity. But still there is a requirement of adding new ideas and functionality. The main target of paper [6] is application of wireless hand gesture controlled robot .In this idea, instructions are transmitted through the palm of person. They used technique of image processing for controlling of device. The paper [4] explain about the implementation and design of gesture controlled

robot by using Flex Sensor, Ultra sonic Sensor, Electronic compass and accelerometer connected to Atmega16 Microcontroller. The research paper [5] describes the Robot, which is controlled by a hand Glove Wirelessly via Bluetooth. The Robot is developed by using the input section consisting of sensor, LCD, Display and a Bluetooth Device and the output section which is consisting of NXT Microcontroller, Motor and Camera. The programming is developed in MATLAB.

III. PROJECT DESCRIPTION

Our whole project is composed of two parts; transmitter part and receiver part. A complete picture of both parts is described as follows:

A. Transmitter Section

Block Diagram of transmitter section is shown in figure 1.

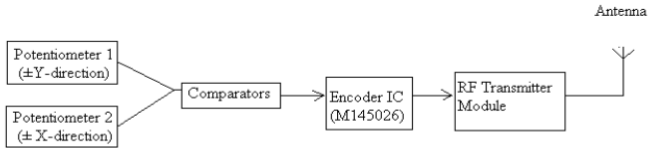


Fig. 1. Block Diagram of Tx Section

Transmitter section of this wireless robot consists of two potentiometers, LM 324 IC (a quad op amp integrated circuit), an encoder IC (M145026) and RF Transmitter (Tx) module as shown in figure 2.

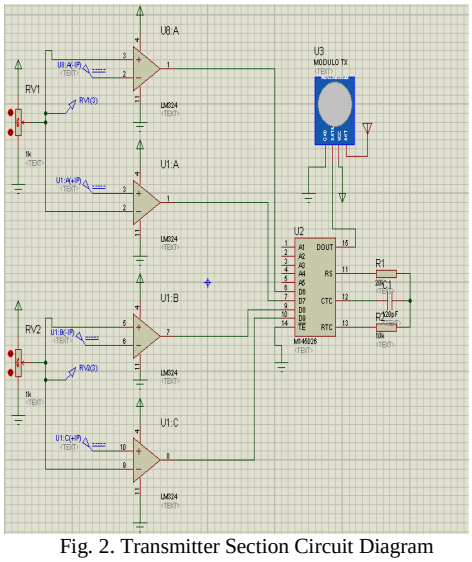


Fig. 2. Transmitter Section Circuit Diagram

Potentiometers

In this project, two potentiometers are used as a voltage regulator. For this purpose, we have set a reference value of 2.5V (50%), at this value of both potentiometer, robot will stay at its static position. When the value of potentiometer 1 increases more than 50%, it has been set that the robot will move in +y-direction similarly, when value is less than 50%, robot will move in opposite direction (-y-direction). In the same manner, the values of potentiometer 2 also have been set for right and left directions.

Comparator

Comparators are usually used to compare the two different levels of voltages and gives high voltage value at the output. In this project, LM 324 IC (a quad op amp integrated circuit) is being used. This IC is more compatible for this project because we have for four different directions that are to be controlled by the PDPs hand gesture, so we need four different combinations of output voltages for each direction. In the comparator circuit, we set fixed value of reference voltage of 2.5V and the comparators give output in the form of digital data.. Therefore we have the following combinations of binary values for each direction that is set by each potentiometer one by one:

- Potentiometer 1 is connected with the first two comparators (comparator 1 and comparator 2) and this potentiometer 1 is used for ±Y-direction. When the value of potentiometer 1 is > 50%, comparator 1 will give output of high logic(logic 1) and of comparator 2 will show low (logic 0), so this combination is used for forward direction. Similarly for reverse or backward direction, the condition will be vice versa.
- Potentiometer 2 is connected with the last two comparators (comparator 3 and comparator 4) and this potentiometer 2 is used for ±X-direction. When the value of potentiometer 2 is > 50%, the output of comparator 3 will be high (logic 1) and the output of comparator 4 will be low (logic 0), so this combination is used for right turn direction. Similarly for left turn direction, the condition will be vice versa.

Encoder IC (M145026)

This IC is used to encode nine bits data and send this data to receiver. It encodes nine inputs with different combinations of codes [9].

The 6, 7, 8 and 9 pins are data pins that are connected with the output of each comparator for receiving the data. The values of pins 11, 12 and 13 are specifically assigned for perfect transmission of data wirelessly [9], while pin 15 is the data out pin which is connected with the data pin of transmitter.

In our project, we have set 362 KHz oscillating frequency for the transmission of data [9].

Proposed Sketch for Hand Gesture

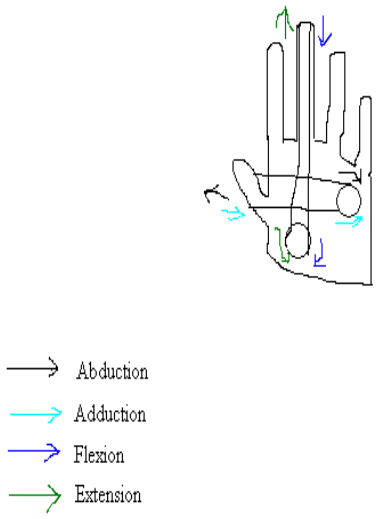


Fig. 3. Tx Section on Proposed Sketch of Hand

In figure 3, it is seen that two potentiometers are placed on the hand of PDPs and both are connected by thumb and middle finger with the help of stick. All four directions of robot are controlled by these movements of hand.

B. Receiver Section

Block Diagram of receiver section is shown in figure 4.

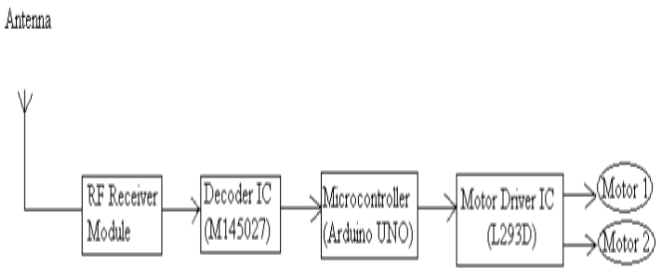


Fig. 4. Block Diagram of Rx Section

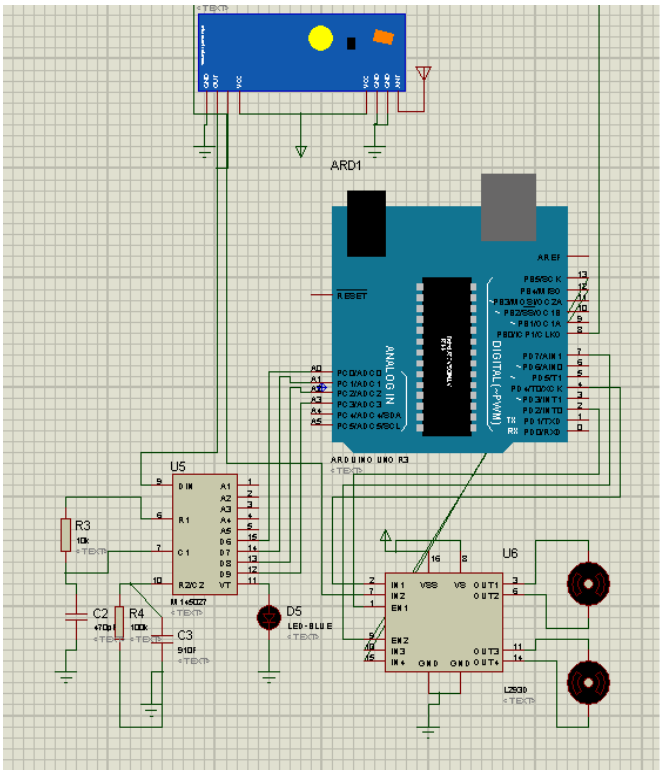


Fig. 5. Receiver Section Circuit Diagram

Receiver section of this wireless robot consists of RF Receiver (Rx), Decoder IC (M145027), L293D Motor Driver IC, Simple DC Motors and Arduino UNO board as shown in figure 5.

Decoder IC (M145027)

Decoders use the same transmitter. It decodes 9-bits data. The M145027 interprets the first five transmitted bits as address and the last four bits as data [9]. The values of pin numbers 6, 7 and 10 are also assigned specifically for the successful transmission of data at 362 KHz oscillating frequency [9]. The pin 9 is the data input pin which receives data from Rx module while 12, 13, 14 and 15 pins are data output pins and are connected with Arduino UNO analog pins (A0, A1, A2 and A3).

Motor Driver IC (L293D)

The L293D is used to control the direction of motors. This IC can be used as bidirectional controlling of 2 motors simultaneously and can also be used as a unidirectional controlling of 4 motors simultaneously. With the help of this IC, we controlled the desired directions of our wireless Robot [10].

The pins 1 and 5 of L293D are connected with pin numbers 2 and 7 of Arduino UNO for controlling the speed of the robot while the input pins of driver IC are also connected with microcontroller to follow the given directions that are generated by hand movements of PDPs.

Arduino Uno

Arduino Uno is a very common microcontroller board build on ATmega328 [3] [11] [1] which has been used for both industrial and academic purposes.This is very flexible to use and very efficient to design desired code.

RF Transmitter and Receiver Module

RF stands for radio frequency [2] [12]. It transmits data of different frequencies according to its model. For encoding and decoding of information we use RF MODULE for data transmission from encoder IC to decoder IC [1].

IV. DESIGN AND WORKING

Design of our robot is composed of two main parts, transmitter part and receiver part. The transmitter part of the project will be incorporated on the hand of PDPs while the robot will contain receiver part of the project. The signal or information of hand movement will be transmitted through RF Module wirelessly and all information of hand movement will be encoded by Encoder IC and then these data will be decoded by Decoder IC.

In our project, we have set 5 conditions that are stop condition, forward movement, and backward movement, movement in right direction and movement in left direction. The speed of all these directions is controlled by PWM (Pulse Width Modulation) technique.

Stop Condition

When the position of the hand is parallel to ground or when palm of the hand is facing upward and both potentiometers have 2.5V, then the output of all four comparators will be high, so at this condition, the robot stops. At this condition, all four LEDs will be switched off (LEDs are attached parallel to DC Motors for clear visualization when simulation runs). You can see all red dots at the outputs of each comparator that indicates the high logic at the output as shown in figure 6.

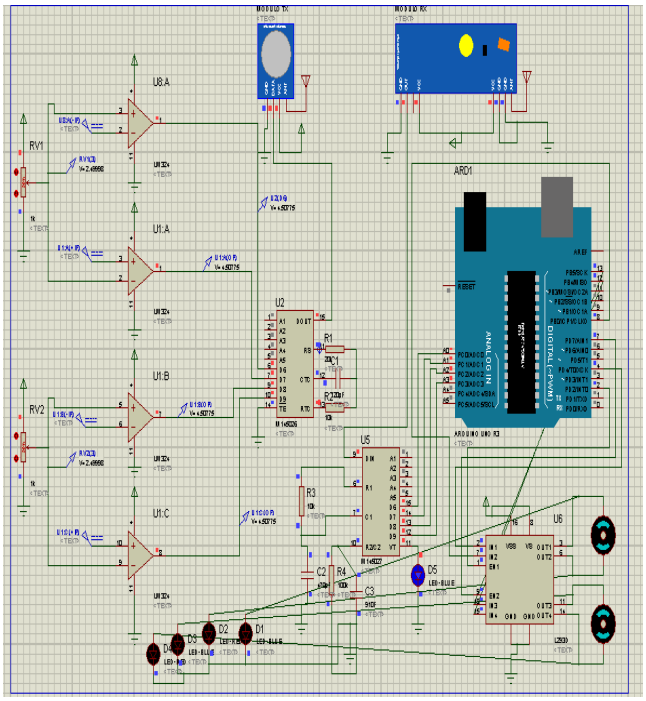


Fig. 6. Stopping Condition

From figure 6, it is clearly shown that both motors are in static condition and all four LEDs that are attached parallel to motors, are also switched off due to following controller commands.

```
digitalWrite(m11, LOW);
digitalWrite(m12, LOW);
digitalWrite(m21, LOW);
digitalWrite(m22, LOW);
```

Above command will only execute, when the output of all four comparators will be high.

Forward Direction

When the hand of the PDPs extends, the robot will move in forward direction. At this condition, the output of first comparator will be high while the output of second comparator will be low. You can easily visualize the red dot at the output of first comparator and blue dot at the second one as shown in figure 7.

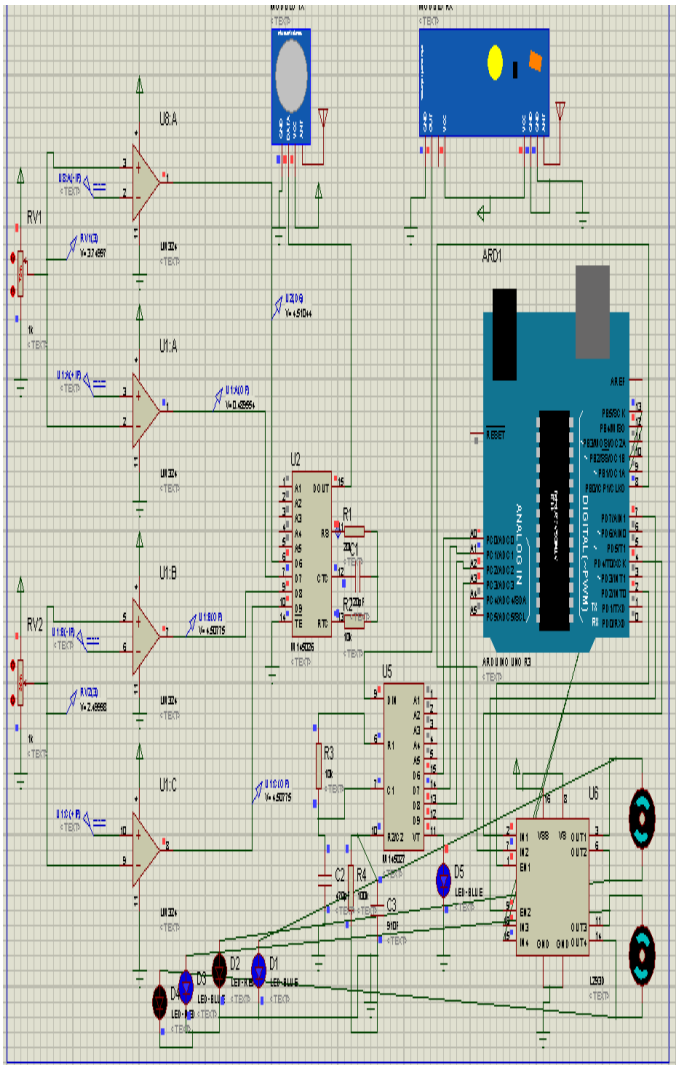


Fig. 7. Forward Movement

From above figure 7, it is shown that both motor move in clock wise direction at its full speed and LED1 and LED 3 are switched on according to the following controller commands:

digitalWrite(m11, HIGH);

digitalWrite(m12, LOW);

digitalWrite(m21, HIGH);

digitalWrite(m22, LOW);

analogWrite(pwm_1,256) ;

analogWrite(pwm_2,256) ;

From above conditions, it is clearly seen that the speed of both motors are set at full speed but in this project, we also control the speed of motor at fixed value for different directions by using PWM technique.

Reverse Direction

When the hand is in the flexion mode, the results of both comparators will be vice versa of Forward direction condition as shown in figure 8.

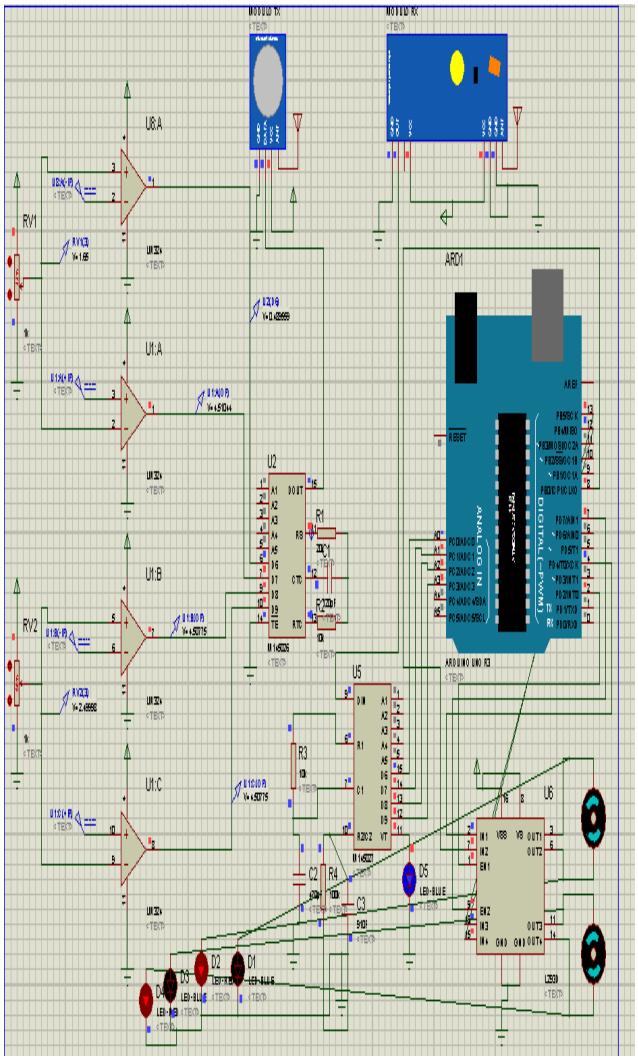


Fig. 8. Backward Movement

From above figure 8, it is shown that both motor move in anti clock wise direction at its full speed and LED2 and LED 4 are switched on according to the following controller commands:

digitalWrite(m11, LOW);

digitalWrite(m12, HIGH);

digitalWrite(m21, LOW);

digitalWrite(m22, HIGH);

analogWrite(pwm_1,256) ;

analogWrite(pwm_2,256) ;

From above conditions, it is clearly seen that the speed of both motors are set at full speed but we can also control the speed of motor at fixed value by using PWM technique.

Right Direction

When the thumb of the hand abducts (moving away from the body), the robots moves in right direction because at this condition, the output of third comparator will be low while the fourth comparator will have high output as shown in figure 9.

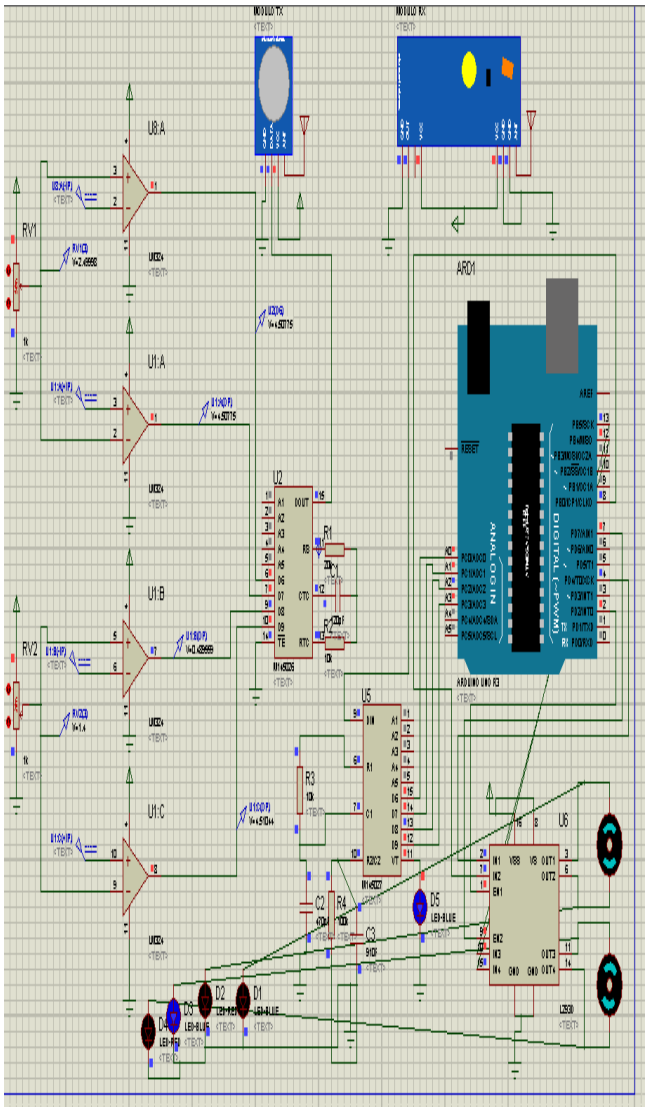


Fig. 9. Right Movement

From above figure 9, it is shown that first motor is in static position while second one moves in clock wise direction at its full speed due to which only LED3 is switched on according to the following controller commands:

digitalWrite(m11, LOW);

digitalWrite(m12, LOW);

digitalWrite(m21, HIGH);

digitalWrite(m22, LOW);

analogWrite(pwm_1,256);

analogWrite(pwm_2,256);

Left Direction

When the thumb of the hand adducts (moving towards body), the output conditions of both comparators will be vice versa of right direction condition as shown in figure 10.

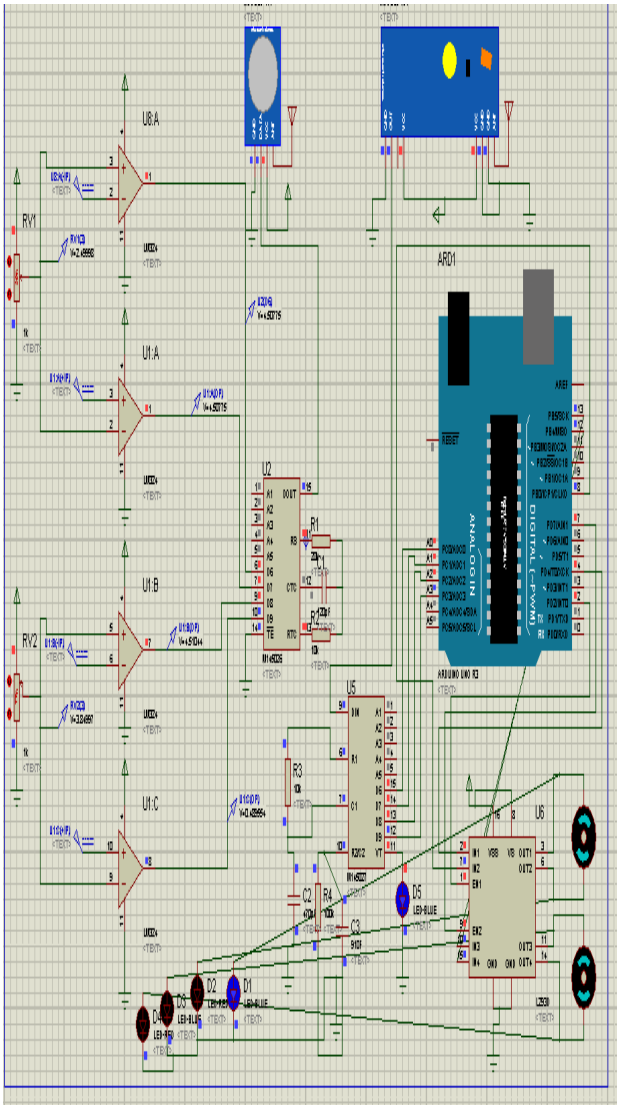


Fig. 10. Left Movement

From above figure 10, it is shown that first motor now moves in clock wise direction at its full speed due to which only LED1 is switched on while second one is in static position according to the following controller commands:

digitalWrite(m11, HIGH);

digitalWrite(m12, LOW);

digitalWrite(m21, LOW);

digitalWrite(m22, LOW);

```
analogWrite(pwm_1,256);
analogWrite(pwm_2,256);
```

V. COMPARISONS WITH EXISTING SYSTEM

The main advantage of our project is the cost effectiveness over the previous designed projects. Previous designed project used accelerometer [1] for ± X and Y directions while we used simple potentiometers that also gave the desired results. Another main advantage and improvisation of our project is the controlling of speed in all directions by using PWM technique.

VI. CONCLUSION

In this paper, an automated wireless control of robot (cost effective robot) has been developed for PDPs which works according to the gesture of the hand. This automatic robot is developed to enhance the independency of the PDPs by controlling the robot through the hand gesture. The presented idea of the undersigned project is taken from literature but it is novel in this sense that this idea is not only a cost effective approach but the speed of the robot can be controlled in any direction.

Our project has some limitation that is our robot has fixed speed controlling value, means once we set the value of PWM, the robot will move in any direction with that fixed PWM value.

The future direction of our designed project is the development of an automated wheel chair for the patients; these robots can also be used in security department by detecting and destroying the bombs.

Following table 1 provides a clear picture of the direction controlling of whole robot.

TABLE I. Summary of Direction Controlling

Sr. No.	Axes	C 1	C 2	C 3	C 4	Robot Directions
1	+ Y	1	0	-	-	Forward
2	- Y	0	1	-	-	Backward
3	+ X	-	-	0	1	Right
4	- X	-	-	1	0	Left
5	-	1	1	1	1	Stop

ACKNOWLEDGMENT

We, the authors of this paper would like to thank Enginner Abdul Haseeb, Lab Engineer of HIET, Hamdard University for his valuable help and direction regarding the development of this robot. We also thank to Dr. Muhammad Faisal Khan, Associate Professor of HIE, Hamdard University and Engineer Rizwan Tanweer, Assistant Professor of HIET, Hamdard University for their valuable time to guide us about writing and citing the research papers. We would also like to thank different open sources and related articles on different components that are used in this project.

REFERENCES

[1] Swarna Prabha Jena, Sworaj Kumar Nayak, Saroj Kumar Sahoo, Sibu Ranjan Sahoo, Saraswata Dash, Sunil Kumar Sahoo "Accelerometer Based Gesture Controlled Robot Using Arduino",

[2] RiyazMansuri, SandeshVakale, AshishShinde, Tanveer Patel, "Hand Gesture Control Robot Vechile", IJECT, Vol-4, Issue SPL-2, PP. 77-80, JAN-MARCH 2013.

[3] Aswath S, Chinmaya Krishna Tilak, Amal Suresh and GaneshaUdupa, "Human Gesture Recognition for Real-Time Control of Humanoid Robot", International Journal of Advances in Mechanical and Automobile Engineering (IJAMAE), Vol- 1, Issue 1, PP.96-100, (2014).

[4] Vicky Somkuwar, RoshanGabhane, Sandeepkakde, "Design and Implementation of Gesture Controlled Robot using Flex sensor and Accelerometer".

[5] GauravGautam, AbhijeetAshish, Anil Kumar, Avdesh, "Wirelessly Hand Glove Operated Robot", International Journal of AdvancedResearch in Electronics and Communication Engineering (IJARECE), Volume-3, Issue-11,PP.-1546-1547, November 2014.

[6] Harish Kumar Kaura, VipulHonrao, SayaliPatil, PravishShetty, "Gesture Controlled Robot using Image Processing".

[7] AdityaPurkayastha, Akhil Devi Prasad, Arunav Bora, Akshaykumar Gupta, Pankaj Singh, "Hand Gestures Controlled Robotic Arm", Journal of International Academic Research For Multidisciplinary, Vol-2, Issue-4, PP.-234-240, May 2014.

[8] Love Aggarwal, Varnika Gaur, PuneetVerma, "Design and Implementation of Wireless Gesture Controlled Robotic Arm with Vision", International Journal of Computer Application, Vol-79, No.-13, PP.-39-43, October 2013.

[9] M145026, M145027 and M145028 Remote Control Encoder/Decoder Circuits Data Sheet (http://www.itisravenna.gov.it/sheet/45027.pdf).

[10] L293x Quadruple Half-H Drivers Data Sheet (http://www.ti.com/lit/ds/symlink/l293.pdf)

[11] Shruthi B. N, Shivraj, Sumathi S, "Hand Gesture Based Direction Control of Robocar using Arduino Microcontroller", International Journal of Recent Technology and Engineering(IJRTE), Volume-3, Issue-3,PP.-32-35, July 2014.

[12] http://oap.sourceforge.net/datasheets/rf.pdf

Industrial Usage of FPGA based Data Acquisition Card (DAQ)

Wajahat Hussain
Design Enginnering Department
KS&EW
Karachi, Pakistan
wajahathussain16@yahoo.com

Maria Andleeb Siddiqui
Dept. of Computer Science and Software Engg.
NEDUET
Karachi, Pakistan
mariasiddiqui88@yahoo.com

Najmi Ghani Haider
Dept. of Computer Science and Software Engg.
NEDUET
Karachi, Pakistan
chaircsit@neduet.edu.pk

***Abstract*—** *The overall productivity and efficiency of the industries can be significantly increased by the usage of Data Acquisition (DAQ) cards in implementation and control functions on digital computers. This paper deals with the design and development of general purpose DAQ card which uses the modern Field Programmable Gate Arrays (FPGAs) technology for the central processor of the system. In virtual instrumentation environment, the FPGA based DAQ provides basic data acquisition functionality for applications such as simple data logging, portable measurements etc. This paper also provides an excellent opportunity to get introduce with the use of FPGA based DAQ card for industrial parameter control. The modular structure and the Graphical User Interface (GUI) of FPGA based DAQ card has also been discussed in this paper.*

Keywords—Computers; Data Acquisition; DAQ; FPGA ;industrial; modular; virtual

I. INTRODUCTION

A. Data Acquisition

For processing, analysis and storage by a computer, processing is required by which the physical phenomenon from the real world are transformed into electrical signals. These electrical signals are converted and measured into digital format. This process is known as Data Acquisition. For processing, it involves the conversion of analog waveform into digital waveform. Data acquisition system includes the folloeing components:

- For conversion of physical parameters to electrical signals, sensors are used.
- To manipulate the analog signals, signal conditioning circuitry is used.
- Conditioned sensor signals are converted to digital values using ADC.

0) Data Acquisition Hardware

Interface between signal and PC is created by DAQ hardware. DAQ hardware could be cards connected to slots (S-100 bus, Apple bus, ISA, MCA, PCI, PCI-E etc.) or it could be connected to computer's port (parallel, serial, USB) etc in the form of a module. ADC, DAC, TTL-IO, high speed timers, RAM) are the multiple components contained in DAQ cards).Microcontrollers are used to access the DAQ card components via bus, that can run small programs.

2) Data Acquisition Software

In order for the DAQ harware to work with PC, DAQ software is needed. Writing and reading on the hardware is performed by low level registers which are enable by device drivers, this is done while a developing user application is exposed to a standard API. There is software COMEDI which allows different operating system to be compatible with some user applications e.g. the user application which runs on BCD and linux will also run on windows.

B. FPGA Technology for Data Acquisition

In the past, FPGAs are used for lower complexity /speed/ volumes designs, nowadays FPGA is capable of easily push the 500 MHz performance barrier. For almost any type of design, FPGAs are a compelling proposition, along with increase in logic density and host of other featuressuch as DSP blocks, clocking, embedded processors and high speed serial at even low price points. Design of data acquisition for process application is presented in [1]. In order to acquire 16 MB real time data, FPGA was configured and programmed in this study. Three parameters of the process plant i.e. temperature, pressure and level is considered for the real time validation and design. It is concluded that the system acquired the online data from different sensors and replaced them. The complex and time consuming steps of floor planning, place and route, time analysis and mask is eliminated by the FPGA design flow. Since the design logic is already synthesized to be placed onto an already verified characterized FPGA device, it respins stages of the product. The use of FPGA technology for the development of an I/O card will offer several advantages such as:

- High speed processing
- Customizable to particular application.
- Concurrent processing as opposed to serial execution in CPU based system.
- Availability of large amount of design resources.
- Field programmability to reconfigure hardware
- Efficient and rapid prototyping testing and debugging

facility etc. Hence the use of FPGA technology in the development of DAQ card enables us to process parameters that have a high frequency response in a more efficient manner. Likewise in [2], serial RS-232 and SPI communication protocol on FPGA platform is used to implement a data acquisition system. Two phases of implementation are used in this paper. In the first phase, SPI interface and in second phase RS-232 interface was implemented. This system results to achieve the data rate of 1.5 Mbps and high accuracy of approximately 99% for the system.

In Application Specific Integrated Circuits (ASICs) where the device is custom built for particular design, unlikely FPGAs have the capability to be programmed according to the desired functionality requirements and applications. SRAM based FPGAs are the most dominant type which can be programmed as the design environment evolve although they are OTP (One Time Programmable)

Advantages to designing with an FPGA (instead of an ASIC or ASSP) include:

- Rapid prototyping
- Shorter time to market
- The ability to re-program in the field for debugging
- Lower NRE costs
- Long product life cycle to mitigate obsolescence risk

A modular 32 channel data acquisition (DAQ) system using National Instruments, PXI chassis, along with FPGA, ADC, Signal generator and timing and synchronization modules was developed in [3]. This system was capable of acquiring input signal frequencies ranging from 100 Hz- 12 MHz. Communication between the system and PC via USB 2.0 was achieved by designing a high speed, high precision, portable data acquisition and processing system. In Altera Corporation cyclone II FPGA platform, Verilog hardware description language was used in this system [4].

A processing system for surface Plasmon resonance (SPR) and a reconfigurable image acquisition was presented in [5]. For sending processed and unprocessed data from a FPGA to a remote device, this system also incorporates communication hardware. By the help of this system, the user is enable to control the strategies of conditioning and signal processing. Similarly [6] presented a comprehensive review on the development of new generation smarter, reconfigurable and lower power consumption sensors based on FPGA in Spain. It concluded that FPGA devices have reached a high level of development that puts them in competition with application specific integrated circuits (ASICs) in terms of performance, power, consumption and costs. The use of data acquisition system in fields of internet, microcomputer, remote data acquisition system and reconfigurable data acquisition system is presented in [7]. For internet based data acquisition system, pressure induced birefringence effects occurred. In micro computers amplification, multiplexing and pre processing is incorporated. In remote, an integrated wireless SCADA was developed and in reconfigurable data, FPGA based system has been successfully integrated with radar optimization. A new sorting control system based on FPGA is proposed in [8]. It concluded that sorting control system could process multi channel and higher – yield signals significantly.

A system with embedded time interval data acquisition system (DAQ) is developed for zero power (ZPR) reactor noise experiments in [9]. The architecture is tested on different FPGA platforms with different speed grades and hardware resources. This is a low cost experimental tool and needs little time to establish. Likewise a system featured with four different channels is presented in [10]. 1 MHz bandwidth, simultaneous acquisition, 9 independent bipolar ranges and maximum sampling rate of 600 Ks/s is used in this study. The usage of these devices significantly useful for advanced measurement systems requiring embedded processing. Sensory circuit based Data Acquisition system is presented in

[11]. The DAQ system collected data is transferred via blue tooth. Web or internet can easily access the data put on server via bluetooth. So controlling the device by client module became easy. FPGA, ADC and Ethernet controller (W12-830 MJ) base data acquisition board is designed and implemented

[12]. The acquired data is displayed using LAB view program and FPGA is used to generate control signals for interfacing between ADC and Network module. Likewise [13] presented the development of FPGA based data acquisition and encryption system. Two stages are incorporated in this design. The first stage is the acquiring of data from real world and its conversion in digital form. The second stage is the encryption and stotrage of data in FPGA memory. Sensory those are mounted on intelligent systems are used to record the data of pressure, temperature, acceleration etc.

1) Features of FPGA

FPGAs - Field Programmable Gate Arrays - are future-oriented building bricks which allow perfect customization of the hardware at an attractive price even in low quantities. FPGA components available today have usable sizes at an acceptable price. The main feature that forms the basis of classification is the interconnection technology of logic cells inside FPGA. The four categories are mentioned below namely

- SRAM – Based
- ANTI FUSED – Based
- EPROM – Based
- EEPROM – Based

TABLE I shows the technology overview of the above mentioned categories.

TABLE I TECHNOLOGY OVERVIEW OF FPGA TECHNOLOGIES

Configuration technology	Technology overview and features
SRAM – based	An external device (non-volatile memory or uP) programs the device on power up. Allows fast reconfiguration. Configuration is non-volatile and cannot be changed.
ANTI FUSED – based	Configuration is set by burning internal fuses to implement the desired functionality. Configuration is non-volatile and cannot be changed.
EPROM – based	Configuration is similar to EPROM devices. Configuration in non-volatile. Device must be configured out of circuit (off board).
EEPROM – based	Configuration is similar to EEPROM devices. Configuration is non-volatile. Device must be configured and reconfigured out of circuit (off board).

II. FPGA DESIGN FLOW

The following steps are consisted in typical FPGA design flow: (some of the steps are combined in advance FPGA design flow)

- Synthesis: RTL code is translated into LUTs, flips flops, memories etc.
- Mapping: LUTs and flip flops are mapped into slices.
- Place and Route: Prior to routing, the placement of all slices, memory etc in the FPGA is first decided and then the routing of the signals that connect these components is performed.
- Bitfile generation: the netlist produces by the place and route step is converted into a bitstream that helps to configure the FPGA.

. Fig. 1 shows the typical FPGA design flow.

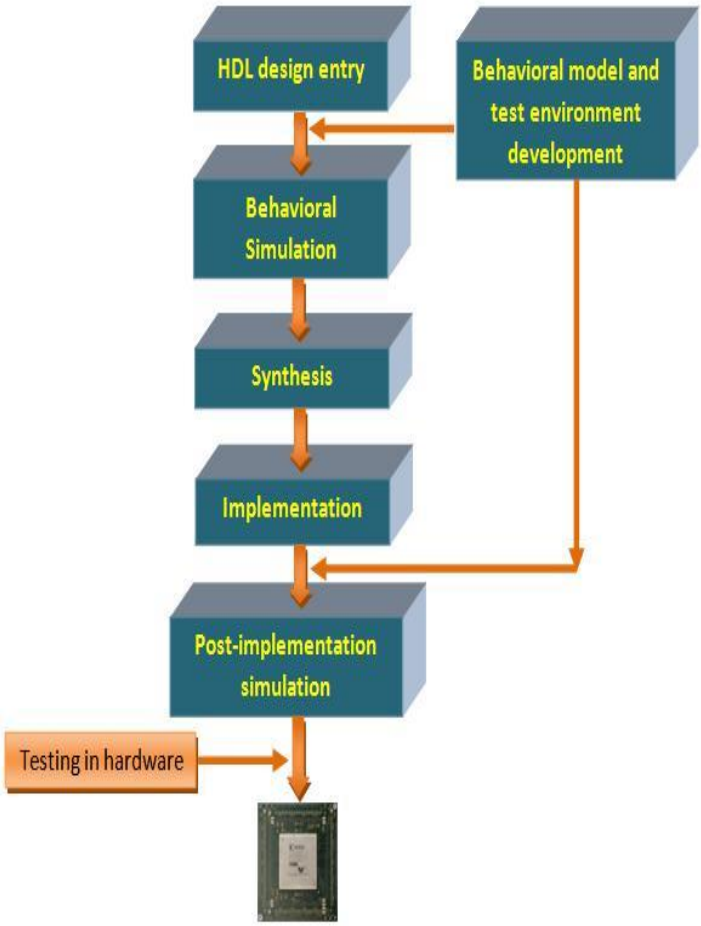


Fig.1. FPGA design Flow

Some steps are optional and only used in certain cases. For e.g. to determine whether the time constraint for the design is meeting or to determine the critical path for certain designs, static timing analyzer is used. But it is not always necessary to use it as if the timing constraint is not met, then the place and route tool will usually print a warning.

A) HDL Design Entry

"Design entry" is referred to involve writing an HDL description. The working of the device is defined by HDL, a class of high level language.

B) Behavioral Model and Test Environment Development

In order to test the HDL design a test design is created to detect errors because it is impossible to develop the fully correct design of HDL at once.Therefore, it should be tested for possible errors. Test environment creation is the software development area; a program is tested by running it. Writing a dedicated test environment is involved in FPGA design.

C) Behavioral Simulation

The HDL description against the corresponding behavioral model is verified by behavioral simulation model. At this stage most of the design errors are fixed.

D) Synthesis

High level HDL description is converted to a machine readable circuit description by an automated process of synthesis. The uncaught errors of behavioral simulation stage can be appeared at this stage, although it is not a problem to synthesize a correctly written HDL code.

E) Implementation

Netlist is converted to an FPGA configuration bitstream by the process known as implementation. It is tailored for specific FPGA design.

F) Post Implementation Simulation

This step is used for the implemented design verification against the behavioral model. Simple designs do not need this step. In this step, for complicated designs, switching and propagation delays are taken into account.

III. METHODOLOGY AND IMPLEMENTATION

This section covers the design and development procedure and HDL architecture of each feature of FPGA DAQ card. Each individual component of the DAQ card is discussed separately with their functional description as well as the designing of their HDL code. Finally complete graphical data flow diagrams of system are presented which further clarifies the architecture.

A) DAQ design Considerations

This section describes the design and development of a computer interfacing system. The main features of the designed system are simple operation, flexible accommodation of various data acquisition and control needs, and adaptability with Internet-based applications. To implement these concepts, the developed system has the following features:

- Interface ability with a PC through its standard port (Serial or Parallel or USB).
- Ability to accept and provide various types of inputs/outputs (analog and digital) with minimum hardware.
- User-friendly graphical user interface (GUI) to monitor and control a data transfer between a PC and

an engineering system.

Future extension capabilities can be introduced through maximizing the use of software in the development.

B) Design of FPGA based DAQ card

We need a means of capturing data from sensors which can be easily used to control our plant. The interface described in this document gives us such a means. It is designed to be easily accessed on windows operating system and from a large number of software applications without the need to write any drivers or additional software.

1) Design Techniques

a) Core Processor

The heart of a DAQ card is Xilinx XC3S500E FPGA which is responsible for input/output and interfacing between PC and outside world. Unfortunately the Spartan FPGAs, are not available separately, they are only shipped in form of the Spartan-3E Starter Kit. The FPGA central processor takes all the digital data from ADC and DI/Os and serially transmits to the computer, where the data is displayed in the LabVIEW based Graphical user Interface (GUI)

b) PC Interface

The available interfacing options are:

- Parallel Port
- Serial Port
- USB Port
- Ethernet Port
- PCI Bus
- ISA Bus

The TABLE II shows the relative data rates of different ports of the computer.

TABLE II. DATA RATES OF DIFFERENT PORTS

Ports	Data rates (Mbps)
Serial Port	0.029
Parallel Port	0.5
USB Port	1.5
Ethernet Port	100

C) Features of FPGA DAQ

From the above mentioned available resources, the solution we selected Xilinx Spartan-3E Starter Kit for the Development of DAQ card.

The following are the main features of this FPGA based DAQ card:

- Xilinx XC3S500E FPGA Based
- Multiplexed 4 Analog input channels
- 1 High Speed Analog Input channel for PID/Filter Implementation
- 8 Digital Input
- 8 Digital Output
- 4 PWM Channels
- Serially plug-in connectivity
- Built-in Wireless Data link for Remote Monitoring
- Data logging Support
- GUI with Installer

In a DAQ process the most important thing is the use of an appropriate DAQ hardware to make a successful interface b/w the signal and the master (PC). Fig. 2 shows the block diagram of FPGA based DAQ card.

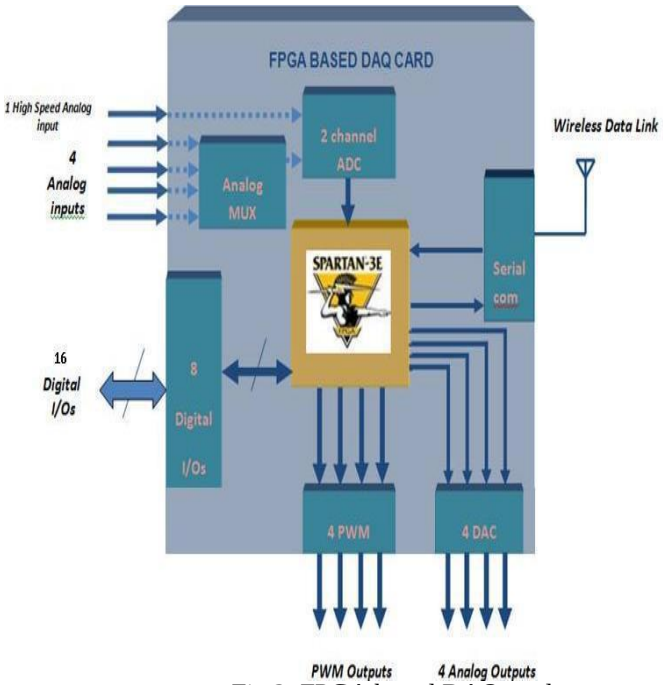


Fig.2. FPGA based DAQ card

1) Multiplexed analog inputs

As mentioned above the built-in ADC is of 2 channels only. In order to increase the channel width, we use external analog multiplexer AD744052 dual 4-1 mux. The selection signal for the mux is carried out by the FPGA. If two or more signals are input simultaneously, then a particular delay is inserted in between for their selection simultaneously. Multiplexing allows several analog signals to be processed by analog-to-digital converter (ADC).

2) Analog Inputs

The signal from sensor is analog, for the further processing we convert the analog signal into digital code. Typically, the input analog (voltage or current) is converted into digital number via ADC, an electronic device. Different coding schemes may be associated with the digital output, such as binary or two's complementary. We have used the built-in LTC1407A-1 ADC this purpose. In Fig 4.2, ADC is a LTC 1407A-1 and the pre-amp is a LTC 6912-1, and they work concurrently when sampling analog data using a serial peripheral interface (SPI). VINA is multiplexed to handle four analog inputs, whereas VIN B is left un-multiplexed for High speed analog inputs. The High Speed Analog Input channel is basically the one of

the channel of ADC LTC1407A-1that is left UN multiplexed. This analog channel supports highest sampling frequency up to 400 KHZ, which make it suitable for implementation of digital PID and other DSP filters.

3) Analog outputs

The signal from the FPGA central processor is digital, for the further processing we convert the digital code into analog. Typically, DAC is an electronic device that converts an input digital code to analog current. We have used built-in LTC2624 for this purpose. The output of DAC served as the analog output of the DAQ card. Fig. 3 shows the digital I/O of FPGA based DAQ card.

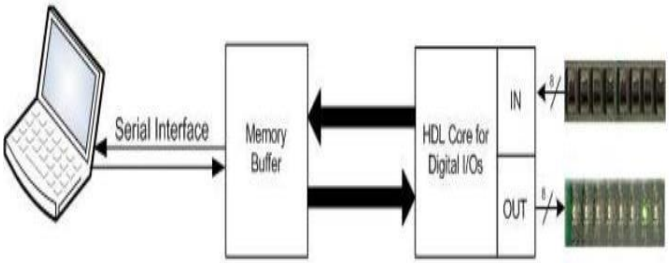


Fig.3. FPGA DAQ with digital I/O

4) PWM channels

FPGA based DAQ card provided with 4 independent PWM channels with frequency of 10 KHz. The Resolution of PWM channel is 8-bit (with a minimum 1% change in duty cycle. Duty cycle is configurable by user interface. To observe the functionality of the PWM outputs, **PWM driver board** is used which fulfills the necessary interfacing requirements between the LVTTL outputs of FPGA and the loads requiring high power. Fig.4 shows the design flow adopted in FPGA based DAQ card.

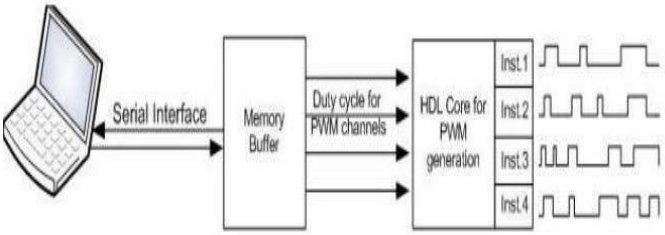


Fig.4. Design flow in designing FPGA based DAQ card

D) GUI FOR FPGA DAQ

On the basis of above discussion, the GUI for FPGA DAQ card has been developed on LabVIEW. The user interacts with the DAQ hardware through an interactive and user friendly GUI which is called Front Panel in the LabVIEW terminology. The front panel configures the DAQ hardware according to user requirements.

Fig.5 shows the labeled illustration of the front panel which shows different features of DAQ card. The sub VIs includes:

- FPGA DAQ Analog Inputs
- FPGA DAQ Analog outputs
- FPGA DAQ Digital Inputs
- FPGA DAQ Digital outputs
- FPGA DAQ PWM outputs
- FPGA DAQ Data logging support

Fig.5 . Front Panel of FPGA based DAQ card.

IV. CONCLUSION

The primary goal of this paper is to explore the modern highly customizable FPGA technology to design and implement an efficient FPGA base DAQ card for industrial design requirements within the limited time frame the entire design was divided into the following tasks:

1. Designing of HDL Core for the Main Processor (on FPGA).
2. Implementing Wireless Communication Protocol.
3. Development of a user friendly GUI with its Installer.
4. Implementation of process control on FPGA Based DAQ.

In future we suggest an External high speed parallel ADC and DAC devices that operate simultaneously instead of the serial operation on the common SPI in the design of the DAQ card for better and more efficient results.

The inputs and outputs can be expanded for future projects as per requirement. A high speed communication interface with the PC can be implemented instead of UART interface. More

efficient FPGA device such as Virtex 5 which has more design resources and offers faster execution speed can be used.

REFERENCES

[1] H.S Murali, M. Meenakshi, "Design and Development of FPGA based Data Acquisition System for Process Automation", Communications in Control Science and Engineering (CCSE), vol.1, pp 7-11, January 2013.

[2] TN Swamy, KM Rashmi, "Data Acquisition System based on FPGA", International Journal of Engineering Research and Applications, vol.3, pp 1504-1509, March-April 2013

[3] S khan, A Borsic, P Manwaring, A Hartov, R Halter, "FPGA based high speed Data Acquisition System for Electrical Impedance Tomography", Journal of Physics:Conference Series, vol. 434, pp 1-4, 2013.

[4] "DSP builder Introduction Handbook", Available: www.altera.com.

[5] EG Pereira, LC oliveria, M Morais, H Neff, "Implementation of a FPGA based data acquisition and processing system for image sensors employed in SPR biosensing", in proceedings of IEEE international Instrumentation and Measurement Technology conference, 12-15 May 2014, Montevideo, Uruguay.

[6] GJ Gareia, CA jara, J Pomares, A Alabdo,LM Poggi, F Torres, "A survey on FPGA based Sensor systems: Towards Intelligent and Reconfigurable low power sensors for computer vision, control and signal processing", Sensors (Basel), vol.14, pp 6247-6278, April 2014.

[7] PJ Bock, S Majumdar, WJ Bock, "Internet Based Distributed Data Acquisition System for Fibre Optic Sensors", IEEE Transaction on instrumentation and Measurement, vol.56, pp 32-38, February 2007.

[8] R Yu, J Cai, H Ma, Y li, "Sorting Control system Design of New Materials: Sorter Based on FPGA", International Journal of Control and Automation, vol.7, pp 253-264, 2014.

[9] M Akrani, H Khalafi, N Vosoughi, "Development of an embedded FPGA Based Data Acquisition system dedicated to zro power Reactor Noise Experiments", Metrology and Measurement Systems, vol.21, pp 433-446, 2014.

[10] PM Pinto, J Gouveia, PM Ramos, "Development, implementation and Characterization of a DSP based data acquisition system with on-board processing", ACTA IMEKO, vol.4, pp 19-25, February 2015.

[11] NA Kulkarni, S.O Dahad, "Intelligent Portable Data Acquisition System with predictive features", vol.3, pp 4103-4110, May 2015.

[12] B Ghetia, A Patel, "Realization of FPGA based Data Acquisition System using Software Embedded Processor and Network Module", International Journal for Scientific Research and Development, vol.3, pp 590-592, 2016.

[13] M Kamtamkar, KS Bapat, "Emedded System based Encryption and Data Acquisition on FPGA for Defense application", International Journal of Advance Research, vol.4, pp 2283-2289, July 2016.

Development of Pre-processing Tools for Urdu Text Processing

Noor ul Ain Ali
Department of Computer Science and Information Technology,
NED University,
Karachi, Pakistan
noorulain.ali89@gmail.com

Sohail Abdul Sattar
Department of Computer Science and Information Technology,
NED University
Karachi, Pakistan
sohail.sattar@neduet.edu.pk

***Abstract*— Pre-processing tools play a considerable role in text processing as they transform the text into structured format, and the data mining algorithms' results are more accurate. In this Paper, the text processing of Urdu documents is implemented by normalizing it with tokenization, stop words filtration, stemming to root forms and multiword normalization. As there is no predefined built-in operator in RapidMiner for Urdu language, so development of these operators has been done.**

Keywords— pre-processing; operators; Urdu; filter stopwords; normalized tokenizer, stemmer, multiwords

I. INTRODUCTION

Urdu is a wide-spoken language in South Asia and Middle East. It is the national language of Pakistan. More than 100 million people speak Urdu. [1] Urdu language comprises of 37 alphabets which are shown in the figure below:

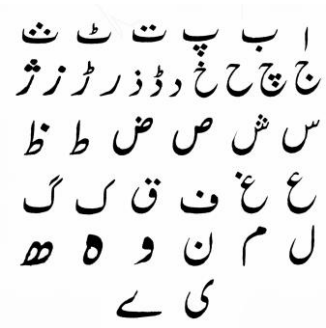


Fig. 1. Urdu language alphabets

Urdu is written in Arabic script in Nastaleeq style from right-to-left. The character set includes basic and secondary letters, aerab, punctuation marks and special symbols. [2] Urdu language supports UTF 8 encoding in various software. Urdu language numerals are shown below:

Fig 2. Urdu language numerals

Urdu computing and text processing started early, in 1980s. It was very difficult to deal with Urdu corpus because of the corpus of Urdu language (i.e. each letter in Urdu language takes different form when placed at different position in a word). [10]

TABLE I. DIFFERENT FORMS OF ب

Different forms of ب	
At the beginning of a word	بھائی
In middle of a word	بلبل
At the end of a word	سرخاب

There is a certain gap for Urdu language text processing and data mining as there are no preprocessing operators developed.

II. PRE-PROCESSING TOOLS FOR URDU

In order to fill the gap for text processing, five pre-processing operators in RapidMiner are developed for Urdu language. The operators are: *Filter Stopwords Operator*, *Normalized Tokenizer*, *Tokenize Multiword Operator*, *Search Multiword Operator*, and *Stemmer*.

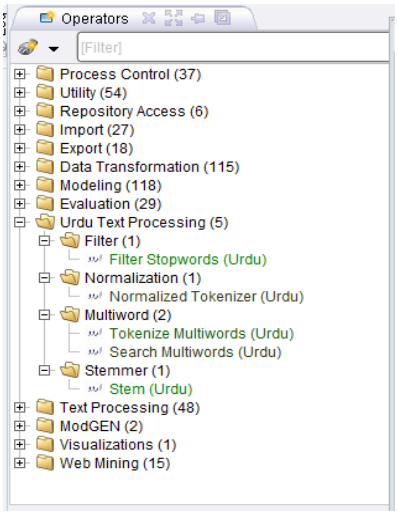


Fig. 3. Urdu Text Processing Extension in RapidMiner

All these operators are based on the text-processing-operators format of RapidMiner, in order to combat compatibility issue.

A. *Filter Stopwords Operator*

Stopwords have no predefined meaning of their own. Every natural language has its own list of stopwords. In information retrieval, filtration is performed to eradicate the stop words from a text document. [4]

In Filter Stopwords (Urdu) operator, a list of stopwords is used that is filtered out from the document. This list is not cent percent accurate but it covers 75% of the stopwords of Urdu language that are commonly used. Some of the stopwords are shown below [8]:

TABLE II. STOPWORDS LIST FOR URDU LANGUAGE

,	و	ہ	ب	أ	ۂ ک
دے	ؤ	ے	ئ	اۃ	ع
:	;	?	!	رے	ای
-	(	)	٨	‘	*

B. *Normalized Tokenizer*

Tokenization is the process of breaking a stream of text into tokens (words, symbols, etc.). The list of tokens is then used as an input for further text preprocessing. The tokenization for Urdu words is basically performed on the spaces. [7]

The function of normalized tokenizer is to remove all the dialectic effects of other languages from Urdu. For example the Arabic ے is replaced with the Urdu ک using this Normalized Operator.

C. *Tokenize Multiword Operator*

The term ‘multi’ means many, and ‘word’ is a single distinct meaningful element of speech or writing. Urdu language has a collection of multiwords that have specific meaning. Some of which are mentioned below [3].

TABLE III. MULTIWORDS LIST FOR URDU LANGUAGE

آگ بگولہ	اثر و رسوخ	غیر ترقی یافتہ
صبح و شام	آمد و رفت	امن و امان

Multiword tokenizer separates all simple and multi words from an Urdu document and displays them in a separate output window.

D. *Search Multiword (Urdu) Operator*

Search Multiword Operator (Urdu) separates only the multiwords from an Urdu document. The multiwords are needed for various purposes; such as to find the frequency of multiwords, etc.

E. *Rule Based Stemmer*

Stemming is a process in information retrieval in which inflected and derived words of any natural language are reduced to their root form. [12]

The process of stemming in Urdu is very complex due to its diverse morphology. [5] Urdu comprises words from Arabic, Persian, Turkish, Sindhi, Hindi, English, and several other languages. Some examples of derived words in Urdu are shown in the table below: [6]

TABLE IV. URDU WORDS DERIVED FROM OTHER LANGUAGES

Derived from language	Words
English	ٹکٹ ڈاکٹر
Arabic	آیت جہاد
Turkish	اردو بیگم
Persian	آشیانہ آسودہ
Hindi	چٹنی پاجامہ

III. RULES USED FOR URDU STEMMER

The following rules are used to develop the algorithm for Rule Based Stemmer for Urdu language. [11]

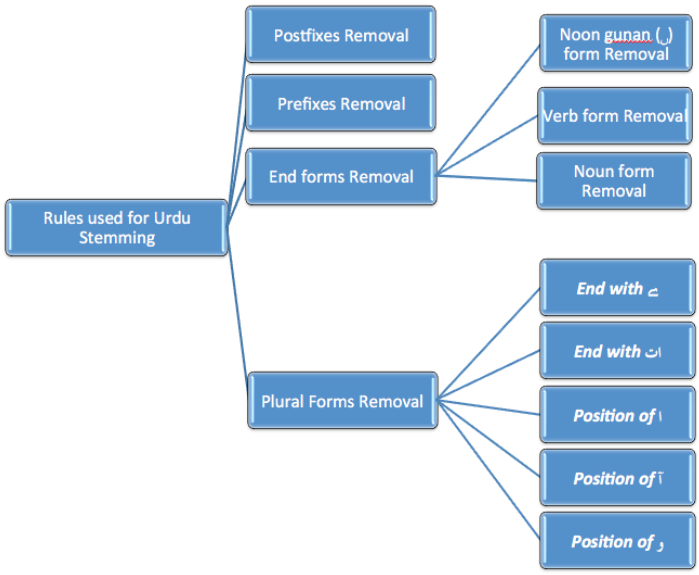


Fig. 4. Rules used for Urdu Stemming

A. *Postfixes*

Postfixes or "لاحقہ" are added at the end of a word to make a new word. Some examples for Urdu language postfixes are given in the table below [3].

TABLE V. LIST OF POSTFIXES IN URDU LANGUAGE

نما	نگاری	نگار	نوس
خم	مندی	مند	آزار

A rule is defined in Urdu stemmer that removes postfixes from the word.

B. *Prefixes*

Prefixes or "سابقہ" are added to the beginning of a word to make a new word. Some of the examples of Urdu language prefixes are [3].

TABLE VI. LIST OF PREFIXES IN URDU LANGUAGE

بن	آشفتہ	آتش	بے	نو
مہ	شہ	صد	بیش	چشم

A rule is defined in Urdu stemmer that removes all prefixes from the word.

C. *End forms*

In Urdu language, a simple word can transform into other noun or verbal form by adding particular suffix at its end. This derivation has been termed as 'End forms'. Generally, the end forms of words and literals in Urdu language are divided into the following main categories:

- Noon gunan (ں) forms
- Verb forms
- Noun forms

1) *Noon Gunan (ں) forms*

Noon gunan (ں) is a special form of ن, which is always present at the end of the word. Some inflected ں forms are shown in the table below:

TABLE VII. LIST OF INFLECTED ں FORMS

اں	ئیں	ویں	یں

2) *Verb forms*

Verb is a word that is used to describe an action, state, or occurrence. In Urdu language, verbs of particular tenses are form by adding predefined suffixes at the end, as shown below:

TABLE VIII. LIST OF INFLECTED VERBS FORMS

یا	تے	تی	تا
ئے	نے	نی	نا

3) *Noun forms*

Noun is a word used to identify any of a class of people, places, or things. In Urdu language, simple noun and verb words are converted into another noun form by adding particular suffixes at the end. Some of these suffixes are:

TABLE IX. LIST OF NOUN SUFFIXES

چہ	گی	پت	پہ

D. *Plural forms*

This is the most complicated part of Urdu language stemming. As Urdu consists of words derived from other languages, therefore there is no particular set of rules that forms plural words. This part has been dealt with as much accuracy as possible, and set a threshold of 80% on the output. Some rules for plural formation have been defined and treated in the stemmer are listed below:

- End with ے
- End with ات
- Position of ا
- Position of آ
- Position of و

1) *End with ے*

This is the most common form of plural formation in Urdu. The last ہ or ا is changed into ے in order to form plural. Some examples of such words are [3].

TABLE X. LIST OF PLURALS ENDING WITH ے

واحد	جمع
لڑکا	لڑکے
دروازہ	دروازے
بستہ	بستے

2) *End with ات*

In this form, ات are added at the end of the word in order to transform it into its respective plural form. Some examples are shown below: [3]

TABLE XI. LIST OF PLURALS ENDING WITH ات

جمع	واحد	الفاظ کی تعداد
آلات	آلہ	3
حالات	حالت	4
روایات	روایت	5
منقولات	منقولہ	6 or more
مطالبات	مطالبہ	

There is one more sub-category in this rule, that is called *"Plural of plural"* or in Urdu "جمع الجمع". Plural of plural is basically the plural form of an already plural word, therefore transforming the word into another level. In order to incorporate this case, this rule cycle is iterated as many times as required. Example of جمع الجمع word is shown below:

TABLE XII. LIST OF جمع الجمع PLURALS

جمع الجمع	جمع	واحد
امورات	امور	امر

3) Position of ا

Several words in Urdu language are transformed into plural forms by adding ا at the 0th and the 3rd index as shown below: [3]

TABLE XIII. LIST OF PLURALS WITH POSITION OF ا

جمع	واحد
اجسام	جسم
اوقات	وقت
اطفال	طفل
اذکار	ذکر

4) Position of آ

Several words in Urdu language are transformed into plural forms by replacing the ا at the 0th index with آ and inserting ا at the 2nd index. Some examples of such words are [3].

TABLE XIV. LIST OF PLURALS WITH POSITION OF آ

جمع	واحد
آثار	اثر
آفاق	افق
آداب	ادب

5) Position of و

Another important category in Urdu morphology is the formation of plurals by adding و at the 2nd index of the original word. Some examples of words that follow this rule are shown below [3].

TABLE XV. LIST OF PLURALS WITH POSITION OF و

جمع	واحد
رقوم	رقم
علوم	علم
سطور	سطر

IV. RESULTS AND DISCUSSIONS

In order to test the pre-processing operators for Urdu, some sample documents have been selected and fed as input in the "Read Document Operator" of RapidMiner.

A. Result of Normalized Tokenizer (Urdu) Operator

When we test the sample document with the default tokenizer of RapidMiner, we get the following output.

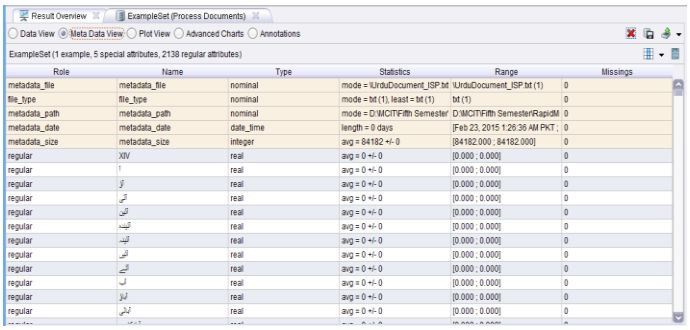


Fig. 5. RapidMiner result of un-normalized (Urdu) words with default Tokenizer Operator

Now we replaced the default tokenizer of RapidMiner with our developed Normalized Tokenizer (Urdu) operator and get the following output (2131 attributes).

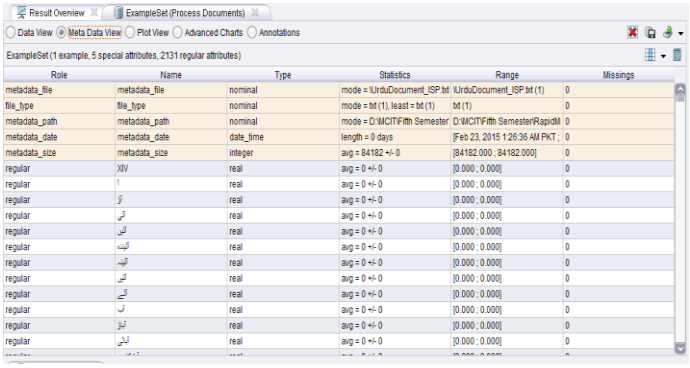


Fig. 6. RapidMiner result of un-normalized (Urdu) words with Normalized Tokenizer (Urdu) Operator

B. Result of Filter Stopwords (Urdu) Operator

When we add the Filter Stopwords (Urdu) operator in the process, we get output as shown in figure below. Note that the total number of tokens after adding the filter stopwords (Urdu) operator has been reduced to 1946 from 2100. It means that the missing 185 tokens were stop words and hence filtered out.

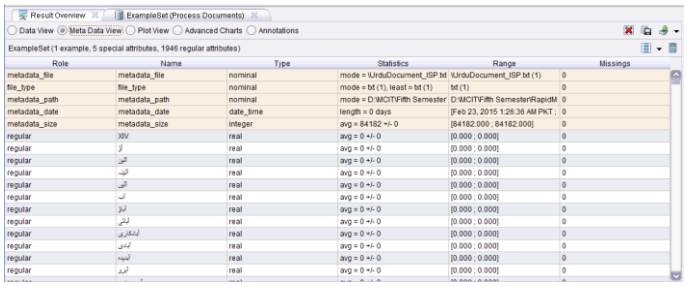


Fig. 7. RapidMiner result of Urdu document with Filter Stopwords (Urdu) Operator

C. *Result of Tokenize Multiword (Urdu) Operator*

When we process the sample multiword document with the default tokenizer present in RapidMiner, we get the following output as shown in figure[#]. As the result shows that all the multiwords are split as singular words, hence their uniqueness being a multiword is disturbed and there is no way to combat that.

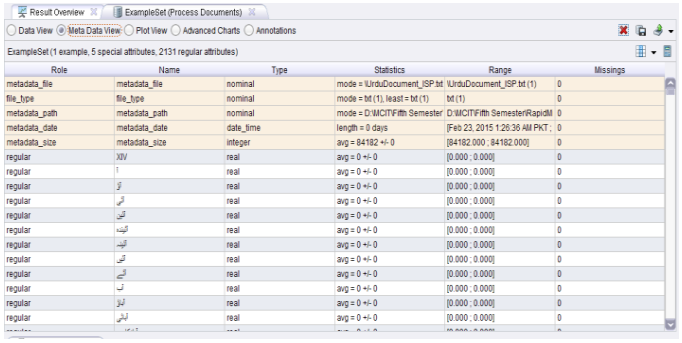


Fig. 8. RapidMiner result of Urdu document without Multiword Tokenizer (Urdu) Operator

By carefully studying the behavior of multiword, we analyzed that there is a need of an operator that retains the uniqueness of multiwords and filters them out from the document. When we add Multiword Tokenizer (Urdu) operator after the tokenizer, we get the output as shown in the figure below. The total number of attributes has increased from 2131 to 2158 because the tokens are shown as separate words and as multiwords both:

Z

Fig. 9. RapidMiner result of Urdu document with Multiword Tokenizer (Urdu) Operator

D. *Result of Search Multiword (Urdu) Operator*

Search multiword operator searches and displays only the multiwords in the document. The output of the operator is shown in the figure below (the total number of attributes is 30 i.e. only Multiwords are displayed):

Fig. 10. RapidMiner result of Urdu document with Search Multiword (Urdu) Operator

The result of Search Multiwords (Urdu) shows that it keeps the uniqueness and grouping of multiword intact. The result can now be exported to any external application, or it can be used as an input for further data mining algorithms like decision tree, graphs, etc.

E. *Result of Stemmer (Urdu) Operator*

When we add Stemmer (Urdu) operator in the process, we get the output as shown in the figure below. Note that the total number of tokens after adding the stemmer has been reduced to 1687 from 1946. It means that the missing 259 tokens were inflected and/or derived words and hence are reduced into their stem forms.

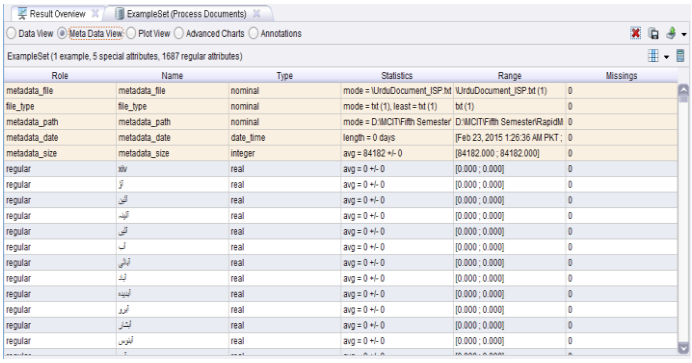


Fig. 11. RapidMiner result of Urdu document with Stemmer (Urdu) Operator

F. *Recall and Precision*

Total number of words in the document: 9502

Total number of (un-normalized) tokens in the document: 2138

TABLE XVI. RECALL AND PRECISION TABLE FOR ALL URDU MORPHOLOGICAL OPERATORS

	Number of relevant records retrieved ***A***	**Total number of relevant records** ***A+B***	**Number of irrelevant and relevant records retrieved** ***A+C***	**Recall** ***A/(A+B) * 100***	**Precision** ***A/(A+C) * 100***
Normalized Tokenizer	7	9	7	77.7%	100%
Search Multiwords	30	44	30	68.18 %	100%
Multiwords Tokenizer	30	44	52	68.18 %	57.69%
Filter Stopwords	185	186	185	99.46 %	100%
Stemmer	259	367	352	70.57 %	73.58%

G. Evaluation of Urdu Text Pre-processing Operators

In the previous section, we calculated recall and precision for all the Urdu pre-processing operators. Precision is also called positive predictive value and is defined as the fraction of retrieved instances that are relevant. Whereas recall is the fraction of relevant instances retrieved and is also known as sensitivity.

The values calculated above shows that the precision for Normalized Tokenizer, Search Multiwords and Filter Stopwords is cent percent, which demonstrates that no false value is retrieved for these operators. The sensitivity for Normalized Tokenizer is 77.7%. In both multiword operators, the recall value is 68.18%, that shows the multiword lists of Urdu is really wide and the multiwords that are not added in the exception list are left out. The Filter Stopwords Operator's filter list is a big list, which contains almost all the stopwords for Urdu language, and therefore the recall for this operator is 99.46% and its precision is 100%.

The stemmer operator for Urdu is a complex operator which contains complex rules for stemming. Some rules defined in the operator creates false positive records and therefore the recall and precision for stemmer is 70.57% and 73.58% respectively.

ACKNOWLEDGMENT

We would like to acknowledge assistance of Prof. Dr. Tafseer Ahmed, Department of Computer Science, DHA Suffa University, Karachi, Pakistan, for his untiring assistance and guidance in this research.

REFERENCES

[1] Sarmad Hussain, "Letter-to-Sound Conversion for Urdu Text-to-Speech System", Centre for Research in Urdu Language Processing, Lahore, Pakistan http://www.aclweb.org/anthology/W04-1613 (visited on September 11, 2016)

[2] S. Hussain, "Letter to Sound Rules for Urdu Text to Speech System", in Proceedings of Workshop on Computational Approaches to Arabic Script-based Languages, COLING 2004, Geneva, Switzerland (2004) http://www.aclweb.org/anthology/W04-1613 (visited on September 26, 2016)

[3] Words collection from Qurat-ul-Ain Akram, Asmar Naseer, Sarmad Hussain, "Assas-Band, an Affix-Exception-List Based Urdu Stemmer", Centre for Language Processing, August 23, 2011; Feroz-Ul-Lughat jame By Maulvi Ferozuddin; and GRE Practical Urdu to English Dictionary (21st century edition) compiled by M. Raza-ul-Haq Badakhshani, Khawaja Ejaz Rasool and M. S. Bhatti (Azhar Publishers, Lahore, Pakistan

[4] Urdu Text Archive http://linguisticslearner.blogspot.com/2011/09/urdu-text-archive.html (visited on September 26, 2016)

[5] Dara Becker and Kashif Riaz, "A study in Urdu corpus construction" http://www.aclweb.org/anthology/W02-1201 (visited on September 29, 2016)

[6] List of English words from Urdu or Hindi Origin https://en.wikipedia.org/wiki/List_of_English_words_of_Hindi_or_Urdu_origin (visitied on September 27, 2016)

[7] Nadir Durrani, "Typology of Urdu word and automated word segmentation in Urdu Text Corpus", August 2007 [MS Thesis, NUCES, Lahore, Pakistan] http://www.cle.org.pk/Publication/theses/2007/typology_of_word.pdf

[8] Stopwords for Urdu language http://www.ranks.nl/stopwords/urdu (visited on September 28, 2016)

[9] Madiha Ijaz, Sarmad Hussain, "Corpus based Urdu Lexicon Development", Centre for Research in Urdu Language Processing, NUCES, Lahore, Pakistan http://www.cle.org.pk/Publication/papers/2007/corpus_based_urdu_lexicon_development.pdf (visited on September 28, 2016)

[10] Agha Ali Raza, Sarmad Hussain, Huda Sarfraz, Inam Ullah and Zahid Sarfraz, "Design and development of phonetically rich Urdu speech corpus", NUCES http://www.cle.org.pk/Publication/papers/2009/Urdu-Speech-Corpus.pdf (visited on September 15, 2016)

[11] Vaishali Gupta, Nisheeth Joshi, Iti Mathur, "Rule Based Stemmer in Urdu", Apashi Institute, Banasthali University, Rajasthan, India, https://arxiv.org/pdf/1310.0581.pdf (visited on September 10, 2016)

[12] Wikipedia: Stemming, http://en.wikipedia.org/wiki/Stemming (visited on September 3, 2016)

Noise Pattern Recognition in Oceanic Environment Using Statistical Characterization of Oceanic Noise in Deep Sea

A Computational Design Approach

Junaid Iqbal
CSSE Department, NED University of Engg: & Tech: Karachi, Pakistan.
engr.junaidiqbal@yahoo.com

Sohail Abdul Sattar
CSSE Department, NED University of Engg: & Tech: Karachi, Pakistan.
sattar@neduet.edu.pk

***Abstract* – Ocean seems to be silent but actually it is not. It is always enriched with a background noise known as ambient noise caused by wind, heat, turbulence and some other activities and natural phenomenon. This noise poses considerable hindrances in undersea acoustic communication which is so essential for Oceanographic Monitoring. Oceanographic monitoring covers numerous undersea applications like oceanographic data collection, pollution monitoring, offshore undersea exploration, detect climate changes, seismic monitoring, disaster prevention etc. With the rolling of years, this noise has been experimentally observed to follow the varying patterns of Gaussian distribution on different widespread geographical regions of the sea. For the design and deployment of any undersea application, it is of utmost importance to determine the variation in the patterns of background undersea noise. This paper has been intended to put forward a computational design approach to compare and find the proximity of experimental/empirical noise pattern with that of hypothesized Gaussian noise pattern using Kolmogorov-Smirnov (*KS*) test statistic. This computation for the noise pattern recognition is so much conducive to simulate the oceanic environment using the hypothesized / controlling parameters of Gaussian noise pattern for the purposes of design and deployment of any undersea application.**

I. Introduction

A growing interest has been found in the monitoring of marine environment in the last few years. The instruments used for this oceanographic monitoring have ranged from small-scale underwater wireless sensor networks to highly complex observation systems [10][11]. Oceanographic monitoring covers numerous undersea applications like oceanographic data collection, pollution monitoring, offshore undersea exploration, detect climate changes, seismic monitoring, disaster prevention etc.

For the remote interrogation of the aquatic environment, acoustic communication is one of the foremost tool to represent the only mechanism for a quick propagation of energy over large distances [9]. But the underwater acoustic communication channels face a considerable hindrances created by the background noise of the region in which the communication system is deployed. Ambient noise is so different in warm/shallow waters than the colder/deep water. Based on ubiquitous observational noise evidences found on different widespread geographical regions of the sea, shallow waters noise can be approximated to a lognormal pattern [3][8] whereas the deep waters noise can be approximated to a Gaussian pattern [2][4].

This paper has been intended to put forward a computational design approach to compare and find the proximity of experimental / empirical noise pattern with that of hypothesized Gaussian noise pattern using Kolmogorov-Smirnov (KS) test statistic. Noise pattern recognition is so much conducive to simulate the oceanic environment using the hypothesized / controlling parameters of Gaussian noise pattern for the purposes of design and deployment of any undersea application.

II. Statistical Characterization Of Oceanic Noise In Deep Sea

Estimating from a large set of data, it describes the underlying energy probability density function (pdf) in a discrete time windowing of appropriate length. The unanimous pattern of pdf acquired from all specified seashores/shallow locations found to follow the lognormal distribution law. Whereas pdf pattern is found to be Gaussian in deep sea [1][2][4].

From a variety of data sources, long-term time series data measurements of underwater ambient noise have been analyzed to get temporal fluctuation statistics. In this way, several sets of long-term time series data measurements of

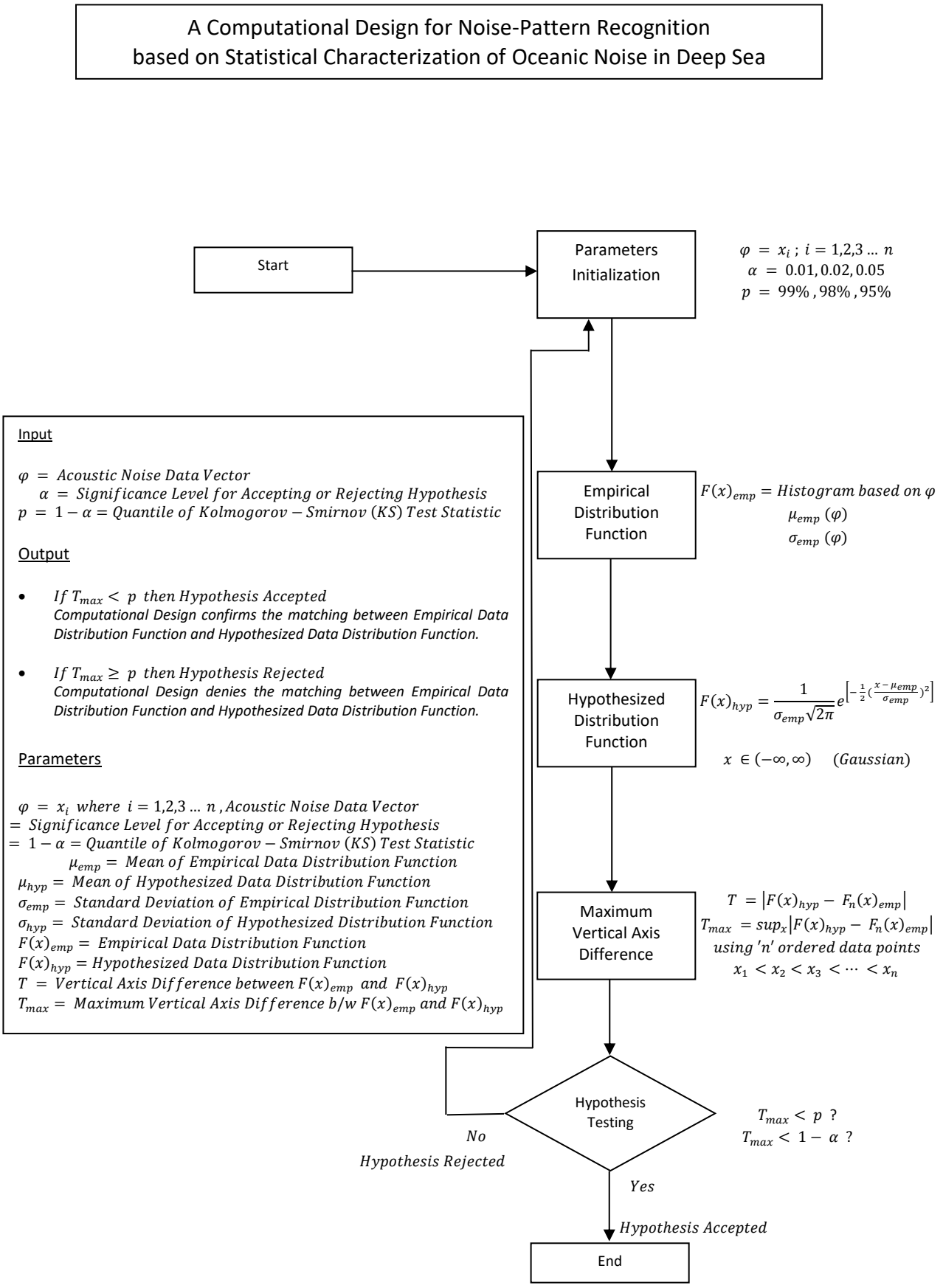


Figure 1: A Computational Design for Noise-Pattern Recognition based on Statistical Characterization of Oceanic Noise in Deep Sea

underwater ambient noise (omnidirectional and beam noise data) from towed sensor arrays and fixed sensors were analyzed in order to identify the core and meaningful statistical characterization parameters [1].

One of the most important and noticeable observation from the time series data [1] yields the typical histograms of noise level (measured in dB) which are closely matched with the Normal (Gaussian) Distribution with a specific mean μ and standard deviation σ. In such cases, the core properties of noise is truly reflected by selecting only two statistical characterization parameters as mean and standard deviation [1].

III. KOLMOGOROV-SMIRNOV (*KS*) TEST STATISTIC

Referencing the Conover (1999), Kolmogorov-Smirnov test statistic (also known as *KS* test statistic) comes under a supremum class of Empirical Distribution Function (EDF) statistics and this statistics has its root or foundation on the maximum vertical-axis difference between the empirical and hypothesized distribution function.

The Kolmogorov-Smirnov one-sample test is a non-parametric alternative to the chi-square goodness-of-fit test. The test compares a cumulative distribution function based on sample observations with some specified population distribution from which the random sample has been drawn. The hypothesis to be tested is that the random sample comes from a completely specified distribution. The Kolmogorov-Smirnov one-sample test is also used for testing hypothesis about discrete distributions. The test is more powerful than the chi-square test.

Given n ordered data points, $x_1 < x_2 < x_3 < \cdots < x_n$, the test statistic is proposed by Kolmogorov (1933) and defined by Conover (1999) as [6][7]:

$$T_{max} = sup_x|F^*(x) - F_n(x)| \quad \text{..... (1)}$$

Here 'sup' refers to supremum which signifies the greatest. $F^*(x)$ denotes Hypothesized Distribution Function and $F_n(x)$ is the Empirical Distribution Function whose estimation is based upon the random sample. In *KS* test statistic of normality, $F^*(x)$ is modeled to be a normal distribution with the known mean μ, and standard deviation σ.

KS test statistic is intended for testing,

$$H_o : F(x) = F^*(x) \; for \; all \; x \; (-\infty \; to \; \infty)$$

(A specified distribution is followed by data).

$$H_a : F(x) \neq F^*(x) \; for \; atleast \; one \; value \; of \; x$$

(A specified distribution is not followed by data).

Referencing the table of quantiles for the *KS* test statistic, if T is exceeded from the $1 - \alpha$ quantile then we reject the Hypothesis H_o at the level of significance α.

IV. A COMPUTATIONAL DESIGN FOR THE NOISE-PATTERN RECOGNITION USING KOLMOGOROV-SMIRNOV (*KS*) TEST STATISTIC

In this section, a computational design has been proposed for the noise-pattern recognition based on the statistical characterization of oceanic noise in the deep sea. This design acquires the acoustical noise data or data samples for a specified time t and allocates it in a variable φ and assumes a hypothesis with a specified level of significance α. It then extracts the $F(x)_{emp}$ EDF (Empirical Distribution Function) along with its mean $\mu_{emp}(\varphi)$ and standard deviation $\sigma_{emp}(\varphi)$ from the data allocated in variable φ as shown in fig: 1. Computational design utilizes the $\mu_{emp}(\varphi)$ and $\sigma_{emp}(\varphi)$ to create a hypothesized PDF (Probability Density Function) $F(x)_{hyp}$ which is Gaussian pattern for the deep sea as mathematically defined below:

$$F(x)_{hyp} = \frac{1}{\sigma_{emp}\sqrt{2\pi}} e^{\left[-\frac{1}{2}\left(\frac{x-\mu_{emp}}{\sigma_{emp}}\right)^2\right]} \quad \text{..... (2)}$$

$$x \in (-\infty, \infty) \quad (Gaussian)$$

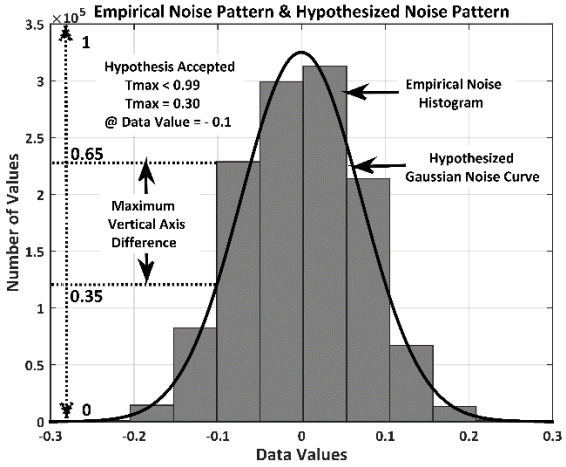


Figure 2: Comparison of Experimental / Empirical Noise Pattern and Hypothesized Gaussian Noise Pattern

Having n ordered data points, $x_1 < x_2 < x_3 < \cdots < x_n$, the computational design calculates the maximum vertical axis difference between $F(x)_{hyp}$ and $F(x)_{emp}$ as given

$$T_{max} = sup_x\left|F(x)_{hyp} - F_n(x)_{emp}\right| \quad \text{..... (3)}$$

The computational design compares T_{max} with $1 - \alpha$. If T_{max} is exceeded from $1 - \alpha$ quantile as shown by the quantile table for the *KS* test statistic, then it rejects the hypothesis regading a hypothesized Probability Density

Function (PDF). Otherwise, it accepts the hypothesis and indicates or recognizes that the statistical pattern of data values allocated in variable φ belong to one of the possible pattern found in the deep sea.

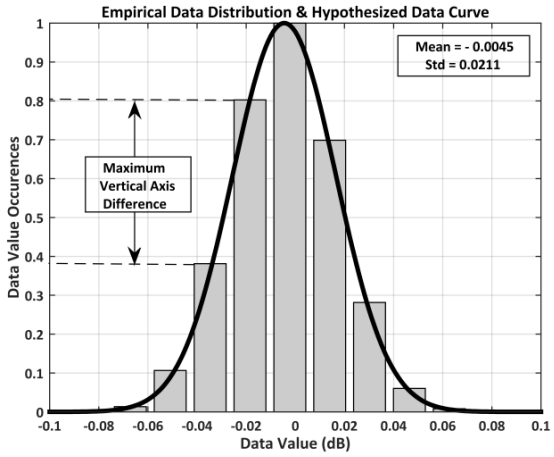


Figure 3: Comparison of Experimental / Empirical Noise Pattern and Hypothesized Gaussian Noise Pattern at $\mu = -0.0045$, $\sigma = 0.0211$.

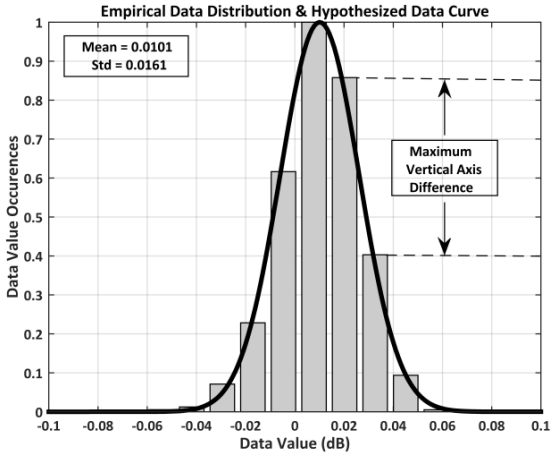


Figure 4: Comparison of Experimental / Empirical Noise Pattern and Hypothesized Gaussian Noise Pattern at $\mu = 0.0101$, $\sigma = 0.0161$.

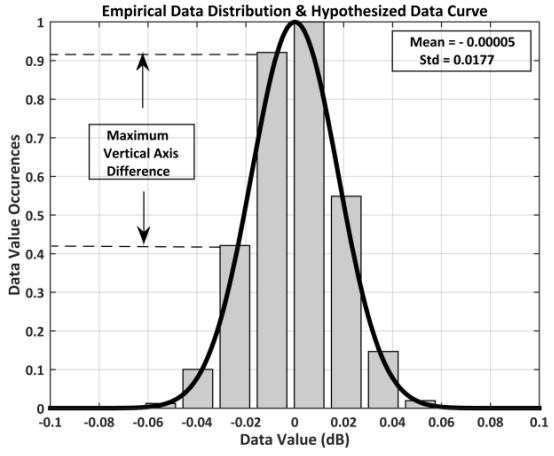


Figure 5: Comparison of Experimental / Empirical Noise Pattern and Hypothesized Gaussian Noise Pattern at $\mu = 0.00005$, $\sigma = 0.0177$.

V. RESULTS

For the testing and verification of the results obtained by the proposed computational design, we referenced the time domain maps or statistics of acoustic noise (dB) from a detailed version of referenced paper mentioned at [1]. Occurrence or Frequency of data values have been mapped on the scale of 0 to 1. Some different empirical data distributions have been chosen based on their different mean and standard deviation which give them slightly a unique shape or pattern shown in fig. 3, fig. 4 and fig. 5.

Utilizing the mean and standard deviation of empirical data distribution, we hypothesized the corresponding Gaussian data curve. The comparison of empirical data distribution against its corresponding Gaussian data curve was performed by using KS test statistic. In all cases, empirical data distribution is closely found to be matched with the corresponding hypothesized Gaussian data curve. These results clearly indicate that the proposed computational design is fully capable of recognizing a diversity of noise patterns found in the deep sea.

VI. SIGNIFICANCE AND IMPLICATION

Dynamic sampling acquired in aquatic environment is one of the most important tasks of the aquatic surveying and undersea exploration. Such sort of surveying and exploration is so much crucially important and is often needed by a diverse range of applications [12]. One of the most recent application is the deployment of Wireless Sensor Networks in aquatic environment. Underwater Wireless Sensor Networks (UWSN) use acoustic communication channels which spatiotemporally face considerable impediments in smooth communication due to the involvement of background ambient noise. In this paper, the proposed computational design performs noise pattern recognition which is so helpful for simulation scenario in designing and pre-deployment phase of WSN. It also opens avenue to suggest such a suitable aquatic regions which offer minimum ambient noise hindrances in acoustic communication.

CONCLUSION

Ubiquitous observational evidences of noise taken at different widespread geographical regions of the deep sea have been found to follow the Gaussian noise pattern. In this paper, a computational design has been proposed and tested to compare the empirical noise data pattern with that of hypothesized Gaussian noise data pattern based on

Kolmogorov-Smirnov (KS) test statistic. The verification of this comparison is in agreement. This computational design pattern in consideration. Additionally, it analyzes the changeability in the pattern of background ambient noise which is so conducive to simulate the oceanic environment for the oceanographic monitoring.

REFERENCES

[1] Anthony I. Eller, "Long Term Temporal Statistics of Ambient Noise", 1st International Conference and Exhibition on Underwater Acoustics Proceedings, 23-28 June 2013, Corfu Island, Greece.

[2] Milica Stajanovic, James Preisig, "Underwater Acoustic Communication Channels: Propagation Models and Statistical Characterization", Underwater Wireless Communication, IEEE Communication Magazine, January 2009.

[3] John R. Potter, Lim Tze Wei, Mandar Chitre, "Ambient Noise Environment in Shallow Tropical Seas and the Implications for Acoustic Sensing", Acoustic Research Laboratory, c/o Physics Department, NUS, 10 Kent Ridge Crescent, Singapore 119260.

[4] Peter H. Dahl, James H. Miller, Douglas H. Cato, Rex K. Andrew, "Underwater Ambient Noise", Acoustics Today, January 2007.

[5] John R. Potter, Lim Tze Wei, Mandar Chitre, "High-Frequency Ambient Noise in Warm Shallow Waters", Acoustic Research Laboratory, c/o Physics Department, NUS, 10 Kent Ridge Crescent, Singapore 119260.

recognizes the noise pattern keeping the Gaussian noise

[6] Nornadiah Mohd Razali, Yap Bee Wah, "Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson Darling Tests", Journal of Statistical Modeling and Analytics, Vol.2 No.1, 21-33, 2011.

[7] M. A. Stephens, "An appreciation of Kolmogorov's 1933 Paper", Technical Report No. 453, June 15, 1992, Department of Statistics, Stanford University, Stanford, California.

[8] Xiaobo Tan, "Autonomous Robotic Fish as Mobile Sensor Platforms: Challenges and Potential Solutions", Marine Technology Society Journal, Biomimetics and Marine Technology, Volume 44, Number 4, July/August 2011.

[9] Kai Chen, Yi Zhou, Jianhua He, "A Localization Scheme for Underwater Wireless Sensor Networks", International Journal of Advanced Science and Technology, Vol. 4, March, 2009.

[10] Jun-Hong Cui, Jiejun Kong, Mario Gerla, Shengli Zhou, "Challenges: Building Scalable Mobile Underwater Wireless Sensor Networks for Aquatic Applications", Computer Science and Engineering Department, University of Connecticut, Storrs, CT 06029.

[11] Cristina Albaladejo, Pedro Sanchez, Andres Iborra, Fulgencio Soto, Juan A. Lopez, Roque Torres, "Wireless Sensor Networks for Oceanographic Monitoring: A Systematic Review", Sensors, ISSN 1424-8220; DSIE, Technical University of Cartagena, Campus Muralla del Mar s/n, Cartagena, E30202, Spain.

[12] Bo Dong, "A Survey of Underwater Wireless Sensor Networks – Localization System Design", Department of Computing Sciences, Texas A&M University-Corpus Christi TX, 78412.

The 1st International Conference on Innovations in Computer Science & Software Engineering, (ICONICS 2016), December 15-16, 2016

IoT Enabled Solution for Monitoring Health of Crops

Ramal Jafri*, Murk Marvi†, Laeeq-uz-Zamanz‡, and Muhammad Khurram§
Computer & Information Systems Engineering, NED University, Pakistan
Email: *ramal.jafri@gmail.com, †marvi.mk15@gmail.com, ‡laeeq.uz.zaman2410@gmail.com, §mkhurram.ned@gmail.com

***Abstract*—Introducing automation in different field of study by exploiting Internet of Things (IoT) technology has become a hot topic nowadays. In this research, a solution has been proposed for monitoring health of crops since automation in agricultural field initially requires monitoring. A centralized approach, without multi-hops, following a star topology has been used. The network constitute of two types of nodes i.e., sensor and gateway node. Gateway node is equipped with GPRS module through data collected from sensor node is transferred to cloud in real-time. Five parameters have been monitored including ambient temperature & humidity, soil moisture, underground temperature, and light intensity. Testbeds has been designed and deployed at two different cites i.e., Civil department NED UET and Gadap Karachi, Pakistan. The results obtained in real time are satisfactory and can be used by researchers for introducing automation in agricultural sector.**

Keywords—Internet of Things (IoT), Cloud, Sensor nodes, Agriculture

I. INTRODUCTION

Internet of Things (IoT) has become a very attractive solution for automation in various fields of study and now days very hot topic in the field of research as well. According to definition given in [1] IoT generally refers to devices i.e., commonly known as sensors, having capability of computing and generating data in real time about different parameters of interest. It will have a great impact on the way we people live and will bring transformation in various ways. For example IoT has started penetrating into homes by introducing intelligent appliance/devices transforming it into a smart home. On the other hand, smart health products are transforming the way urgent and at distant, emergency situations can be handled in a proper way since field of medical is always very critical to even minor issues. Therefore, IoT has great impact in the field of health and sciences as well, helping people to have access to solutions in real time no matter wherever they are. Apart from mentioned applications, IoT has potential to transform greatly various other fields of study i.e., Agricultural, Energy monitoring and production, transport traffic control, and others related. Although IoT brings lots of benefits but there remains few open research challenges that need to be solved for complete realization of this cutting edge technology. Few main hurdles in the vast deployment of IoT enabled devices are cost, network/interoperability, security, reliability, and power consumption. Various industries and research organizations have predicted the growth and need of this technology in near future. According to Cisco in [2], it has been expected that by 2019, more than 24 *billion* IoT connected objects will be deployed. On the other hand, 100 *billion* products will penetrate the market by 2025 according to Huawei forecasts [3]. It has also been predicted by McKinsey Global Institute in [4] that IoT impact on global economy will be around $1.11 to $3.9 by 2025.

Agricultural, on the other hand, is one of the most important sectors for economical growth and plays crucial role for socioeconomic growth of any developing country. Pakistan, by birth received most of the land rich for cultivation therefore contributes a lot in development of our country. However, a decline has been expected in this sector due to lack of natural resources and global warming, causing severe hurdles in its growth. Such problems cannot be solved through traditional manual approaches, being used since decades, because of the involved complexity. For handling such situations, optimal solutions are required which take under considerations all minor and major factors through detailed analysis. Since decades, technology has replaced traditional approaches with computationally effective and intelligent ones for solving such complex problems. Therefore, in this research work we have tried to focus on the problems faced in agricultural sector by providing IoT enabled solution for monitoring farms so that their quality and quantity can be improved.

A. Related work & Motivation

In literature various solutions have been presented which specifically focus on the problem of Agricultural sector. Mainly available solutions can be classified into two broad categories based on their topology i.e., centralized or distributed. Centralized approach is based on a local wireless sensor network having a gateway node through which data is transferred to server i.e., local or global. While in distributed approach each sensor node has direct connectivity to main server. Most of the solutions available in literature for monitoring health of crops are limited to only local wireless sensor networks. In [5] a survey of wireless sensor networks has been presented specifically focusing on Zigbee as a wireless protocol. Although Zigbee is a reliable and power efficient protocol however the radio modules which integrate this protocol are very expensive. Since IoT has started emerging and people are more interested in having everything, at their hand, in real time. Therefore existing solutions limited to local networks will not capture the attention and hence cannot sustain for long.
In [6], the authors have presented a centralized solution with multi-hops. At maximum four sensors can be interfaced with their sensor node including underground temperature, soil moisture, CO_2, ambient temperature & humidity sensor. They have used ATmega1281 micro-controller and AT86RF230 radio module which are expensive as compared to the ones

used in this research. In addition to it, there gateway node is powered through mains with certain back up plans. Therefore additional relay nodes have been installed in between so that the data from sensor nodes can be relayed to gateway node. The design proposed in [6] is clearly very expensive and complex as well. Although, there sensor nodes consume less power as compared to ones proposed in this research however the overall utilization will remain same due to deployment of additional relay nodes in [6]. As opposed to centralized approach, in [7] a distributed approach has been used and authors have performed both monitoring and actuation of various crops including cucumber, brinjal, chilies, tomato, and papaya. Although, they have provided in depth analysis of results but there design is very bulky and heavy in size. There system gets powered through a main which is not really implementable in open fields where such resources are either limited or not available.

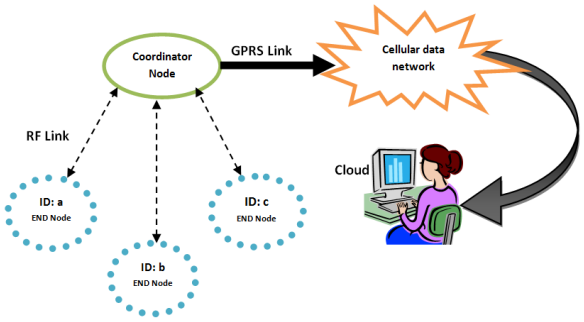


Fig. 1. General model of the proposed solution

In [8] a detailed survey on the challenges and issues in wireless sensor networks, which can be extended to IoT, has been presented. Main challenges mentioned in [8] are production cost, operating environment, network topology, transmission media, fault tolerance, scalability, and power consumption. The proposed solution in this research considers almost all the mentioned issues in [8], as explained in section II and III. Precision Agricultural Monitoring System (PAMS) has been proposed in [9]. The authors have used a fixed NPUMote with two dry cells battery and tried to use power efficiently by exploiting different optimization techniques. There complete model consists of three types of nodes i.e., sensor, root, and gateway. Sensor nodes transfer their data to root nodes from where the data is transferred to gateway node. A central web server has been designed by authors in [9] where data transferred from gateway has been stored. It must be clear that approach present in [9] is complex and expensive as well due to additional root nodes. Apart from that, authors have not mentioned specifically about the selection of sensors and results are also not presented appropriately. The prime motivation behind working on this research project is its importance, need, and requirement which are expected to grow in near future. The actual challenges that are expected to rise in future are scarcity of natural resources and rapid changes in weather conditions. These mentioned challenges indirectly affect the agricultural sector greatly because of its dependence on these mentioned factors. Since one cannot control natural things however through research we can try to come up with solutions that can help farmers to improve the quality and quantity of food. This sector does not only have effect on the financial outcome of farmer and economy of country. However, it indirectly has a noticeable effect on the life of common people as well. The population is increasing with time and hence the requirement of food products as well. In such situations, if quantity of food will start falling down with time then it will become extremely difficult for common people to survive. Although various solutions have been presented in literature but they are very expensive and farmers especially in Pakistan cannot afford it. Therefore, this research aims at design and development of cost effective solution for monitoring different environmental and soil parameters of crop by exploiting state of the art technology.

B. Contribution & Outcomes

The main contribution of this research are listed as under,

Power efficient solution: Extremely power efficient design has been proposed in this research by optimizing both hardware and software modules of the project. In software, power efficiency have been achieved by enabling sleeping modes that consume least power while in hardware each interfaced sensor, radio module, and heat sink have been controlled through transistor. Therefore, most of the interfaced components are turned on only when their function is required otherwise they are kept off. In section II-A3 detailed analysis of power consumption has been presented.

Real time monitoring: Since IoT is not yet fully realizable due to lack of communication models to support huge number of tiny devices with diverse set of requirements. Therefore, in this research though individual sensor nodes are not directly connected to internet however they all are assigned unique IDs and constitute a wireless sensor network (WSN). Through a gateway node, having global connectivity, the data from individual node is collected and then transferred to cloud in real time.

Cost effective design: Cost is one of the main factors limiting wide scale deployment of IoT enabled solutions especially for under developed countries like Pakistan. Looking at the financial, economical, and other problems of farmers, the authors have tried to provide cost effective solution so that everyone can afford and take maximal benefit out of it for improving their crop productivity.

Reliability: Since IoT enabled devices are supposed to be deployed in environments that are harsh and continuously varying. Therefore it is really important to ensure the reliability of nodes, especially when they are deployed in outdoor, which is the case under consideration here. In this research work, the authors have tried their best to ensure reliability of sensor nodes through testing in different environments and confirmed their smooth operation by removing all bugs and issues.

Light weight solution: In most of existing solutions open source operating systems i.e., TinyOS, Ti-RTOS and others have been used for power efficient and reliable operation of sensor nodes. However, in this work the authors have optimized the code structure and designed a very simple

protocol for local communication without use of any operating system. Therefore, the overall software design is extremely light weighted and hence ensuring performance comparable to solutions with operating systems.

The main outcomes of this research are, PCB of sensor and gateway node which is extremely efficient in design and operation. Software, that has light weight and is much optimized in design. Testbeds of proposed solution deployed at two different places i.e., civil department, NED UET & Gadap, Karachi, Pakistan. Database, consisting of real time readings fetched from deployed testbeds, which can be exploited in future research by using state of the art machine learning techniques for introducing automation in agricultural sector.

II. System Model

In this research work centralized approach has been adopted without any multi-hop. There are basically two main modules i.e., sensor node and gateway node as shown in Fig. 1. The sensor nodes get connected to gateway node following a star topology i.e., direct link from sensor to gateway node and no link among sensor nodes.

A. Sensor Node Design

The main components of sensor node are micro-controller, radio modules, various sensors, and power management circuitry. In this research, the authors have used Atmels Atmega328p micro-controller as computing device. It is an 8-bit microcontroller based on AVR RISC architecture and has 32 *KB* of code memory with a processing speed of 20 *MHz*. The 6 channel ADC with 10*bit* resolution is also available in this microcontroller. It has operating voltage range of 1.8-5.5*V* and five software selectable power saving modes to make system power efficient. It consists of 23 general purpose I/O lines, 32 general purpose working registers, byte-oriented two wire serial interface, programmable serial USARTs, SPI serial interface, and programmable Watchdog timer with on chip oscillator. These key features make it suitable to be used in our application.

1) Radio modules: For local communication between sensor and gateway node *RF* 433 *MHz* Tx/Rx pair has been used. It is the simplest and cost effective module for local communication since it does not implement any protocol. It uses Frequency Shift Keying (FSK) for communication over channel and does not packetize the data in any way. Though there are many other transceivers available in the market i.e., Xbee, Bluetooth, and Wi-Fi having a built in communication protocol but keeping the cost constraints and small data size in mind this cheaper module has been chosen. The *RF* Tx/Rx operates at 5*V* and the range of module is about 20*cm*. However, it can be increased by interfacing an external antenna to Tx/Rx pair. Therefore, Steel Brass antennas with an unstressed length of 16*cm* and total length of 62*cm* have been used for extending the range of transceiver up to certain meters.

2) Interfaced sensors: Number of sensor have been interfaced to micro-controller for measuring different environmental and soil parameters including ambient temperature & humidity, moisture of soil, underground temperature, and light intensity. Ambient temperature and humidity sensor has been chosen because photosynthesis and transpiration are two essential processes necessary for any plant growth and these can easily be influenced by temperature variations. Every plant has its own temperature range and going beyond that range can damage plant growth. Apart from that variation in humidity can also disturb the rate of transpiration and hence can cause most of plant diseases. With the help of moisture analysis one can efficiently use the natural water resource that has expected to decay with respect to time. Therefore in this research, soil moisture sensors at two different levels of plants soil have been used for in depth analysis of soil conditions. Since temperature can greatly affect the moisture of soil hence can get dry quickly therefore underground temperature sensor has also been used at two different levels of plant soil. This, in depth monitoring can help in analysis of how variation in temperature can greatly affect soil conditions.

TABLE I POWER ANALYSIS OF SENSOR NODE AS COMPONENTS TURNED ON

S. No	Condition	Current Drawn
1	by PCB without controller	5.7 *mA*
2	when RF modules turned on	31.5 *mA*
3	when all sensors turned on	42 *mA*
4	when fan turned on	46 *mA*
5	in sleeping mode	< 5 *mA*

For measuring moisture of soil YL69 sensor has been used because it is extremely cost effective and simple to use. This sensor produce analog signal which varies in proportion to the water content of soil. When there is no water it produces maximum signal and as water content increases its signal strength decreases. Therefore, for measuring moisture content the authors used simple formula as given in eq. (1). It operates at 5*V* and can be interfaced to any ADC pin of micro-controller.

$$SM_{content} = (1 - \frac{ADC_{in}}{ADC_{ref}})_v * 100\% \quad (1)$$

For ambient temperature and humidity DHT11 sensor has been used which is also very cost effective. It sends a digital signal on the data pin and uses single wire serial interface. It has an operating voltage of 3.3 - 5*V*. Humidity can be measured up to 20% to 90% with a precision sensitivity of 5% and temperature range up to 0 to +50°*C* and sensitivity of 2°*C*. For measuring underground temperature, sealed waterproof probe version of DS18b20 has been used. It is a one wire digital interface sensor and can measure temperatures ranging from -55°*C* to +125°*C* with 0.5°*C* accuracy from -10°*C* to +85°*C*. Simple Light dependent resister (LDR) has been used to measure light intensity. It works on the principle of photo conductivity, as the intensity of light increases its resistance decreases. Thus, voltage variations can be measured using ADC pin of microcontroller. The ADC data is again converted into percentage by using formula as given in eq. (2).

$$LDR_{intensity} = (\frac{ADC_{in}}{ADC_{ref}})_v * 100\% \quad (2)$$

3) Power management: For ensuring long battery life and self optimization of sensor nodes rechargeable batteries have

been used for required power. The batteries get powered through solar panel interfaced and controlled through a single transistor connected to micro-controller. Each sensor node is attached to a solar panel of 10*watt* with 17.5*Vt* and powered by Lead acid battery with rating 8.4*V*, 800*mA/hour*. With the help of LDR sensor the transistor connecting solar panel and battery has been controlled. Whenever the LDR reading is above certain threshold ($LDR_{intensity} > T$) i.e., showing enough sun light and battery is not enough charged ($charging < T_1$) then the transistor is turned on and battery starts charging. For sustaining reliability in cloudy conditions i.e., no sun light available for few days, the authors have defined a threshold for batteries (T_2). If the voltage of batteries falls below that threshold ($charging < T_2$) and there is not enough sun light then the sensor node will go back to sleeping mode without transferring communicating with gateway node. This back up plan has been defined in software so that batteries does not get drained completely which is a worse situation.

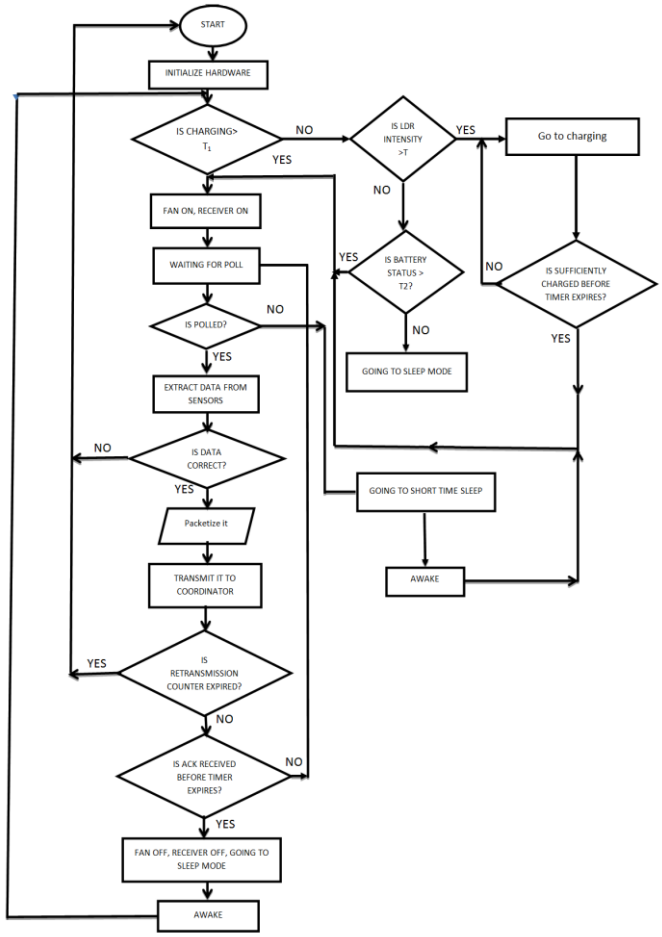


Fig. 2. Flow chart showing functionality of sensor node

For efficient power utilization and management different techniques have been used both in hardware and software design. Power to all interfaced peripherals is controlled through transistors. The ground pin of every component is connected on transistors collector and is controlled by giving high or low signals from micro-controller to its base. This simple technique enabled components to draw current only when they are being used otherwise they are kept off. All sensors have been controlled using single transistor, whenever sensor node needs to fetch data only then all the sensors are turned on. Similarly, the receiver and transmitter are turned on only when they have to transmit or receive the data. At the sensor node a fan has also been interfaced as a heat sink and it is turned on when the node is performing some operation otherwise kept off. For further power optimization, sleeping mode has been enabled through software. Power down mode has been used out of the five available modes since it consumes least amount of power. It keeps most part of the micro-controller off when not in use except WDT, Interrupts and TWI. After successful local communication each sensor node goes to sleep for few minutes i.e., as per defined in software. The authors have carried out in depth analysis of power consumption of each component and complete PCB under certain conditions please, refer to Table. I. It must be clear from Table. I, that most part of the time power consumed by sensor node is less than 5*mA* and at maximum it can reach to 46*mA* which lasts just for few seconds.

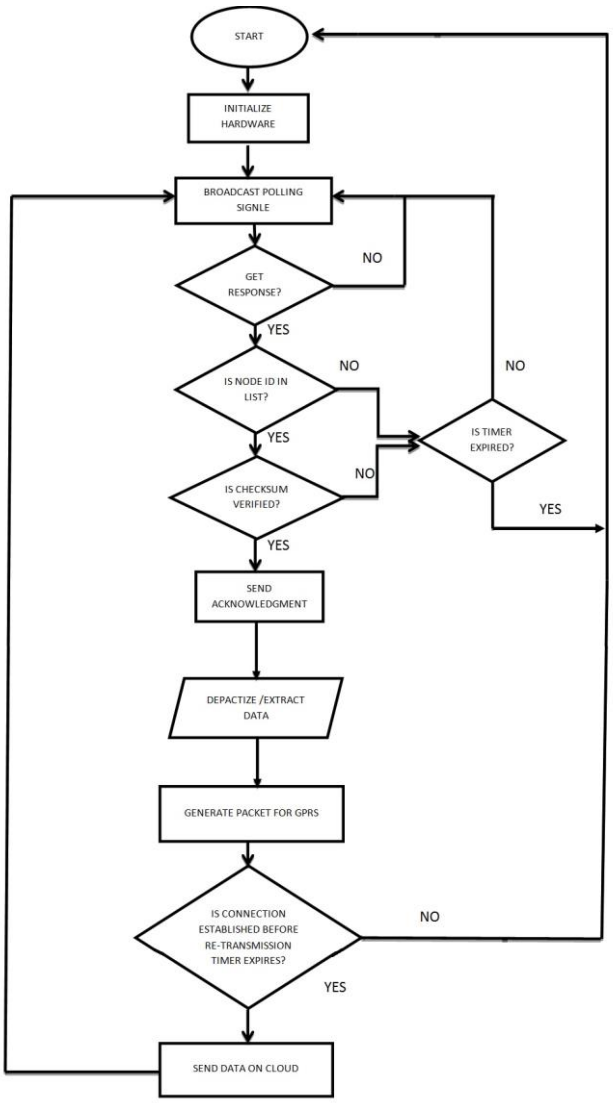


Fig. 3. Flow chart showing functionality of gateway node

B. Gateway Node Design

The main components of gateway node are micro-controller, RF modules, GPRS module, and Power management unit. A fan has also been interfaced as a heat sink which is controlled through transistor as well. GPRS module has also been controlled through transistor because it consumes a lot of power. The RF receiver remains on at gateway node for providing reliable channel to sensor nodes. However, the RF transmitter is controlled through transistor. The microcontroller and RF modules used are same as for sensor node described in section II-A & section II-A1 respectively. Two different prototypes of gateway node have been designed which differ in terms of GPRS module used and power management unit. First prototype of gateway is limited in features therefore cannot be deployed in open fields where no electricity is available and less cellular network coverage as well. However, it outperforms in test fields which does not pose any certain limitations.

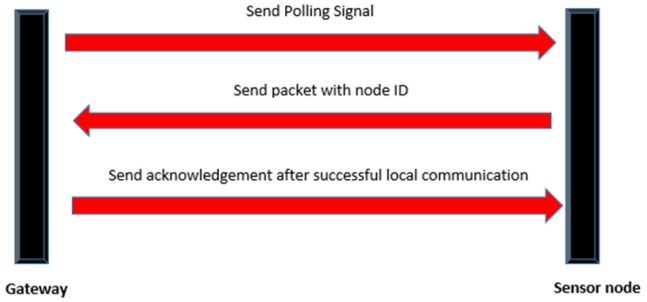


Fig. 4. Local communication flow between sensor and gateway node

1) Gateway node-1: In this initial prototype, Sim800L has been used for global communication which operates in the range of 3.7 to 4.2*V*. Its power circuitry was designed by authors but later on issues of automatic reset and over voltage were faced. After a lot of analysis it was found that the power supply was unstable which was resolved by increasing the input and output capacitance. The discrete Sim800L had a smaller spiral antenna which gave poor performance after casing. Therefore, another antenna was designed by authors using simple wire technique and interfaced with Sim800L, which results in better performance. The Gateway node consumes more power as compared to sensor node since it is continuously on and also has a GPRS that drives 2*A* current while transferring data to cloud. Thus in initial prototype, 12*V*, 2.5*A* adapter connected to main power line has been used for providing power to it. This prototype of gateway node has been installed at civil department of NED University.

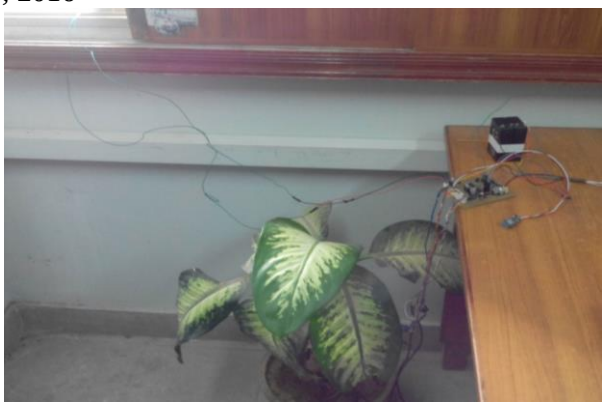


Fig. 5. Testing of nodes under laboratory setup

2) Gateway node-2: In modified prototype, Sim800 module has been used which comes along with power circuitry. It has an SMA antenna as well that does perfect transmission even after casing of the node. The module does not require MAX232 but can directly communicate with controller using TTL serial interface. Simple AT commands are sent by controller to the module in order to establish connection and send data to cloud using TCP connection over 2*G* cellular technology. Different SIM card can be used on the basis of signal strength of a particular network in specific area. This gateway node gets powered through a 12*V*, 5*A* rechargeable battery which gets charged through 20*watt* 25*V* solar panel. A complete separate PCB has been designed for solar panel circuit since there is not any LDR sensor interfaced at gateway node. The battery automatically starts charging when its voltage drop below 12*V* and there is sufficient sun light. A zener diode of 12*V* has been used for controlling the battery charge over time. This prototype of gateway module has been deployed at Gadap, Karachi because electricity is not available there.

TABLE II
PACKET FORMATION FOR LOCAL COMMUNICATION

Field Name	Length (Bytes)	Description
Node ID	1	Start of packet with a unique ID
Payload	35	Data fetched from all sensor
Checksum	7	For error detection
Trailer	1	End of packet

III. Protocol Design and Software Flow

As already discussed in section II-A1, that *RF* module used does not employ any protocol for communication. Therefore, for reliable communication between sensor and gateway node a simple protocol has been designed please refer to Fig. 4. Each sensor node wakes up after certain interval of time and checks its battery status. If it is not sufficiently charge then there are two possible conditions; 1) if enough sun light then it will go in charging mode for certain duration after that waits for getting polled, 2) if sun light is not enough then it will not wait for getting polled and goes back to sleeping mode. If sufficiently charge then a counter will start and waits for gateway to poll till timer expires. If it does not get polled by gateway before timer expire then moves back to sleeping mode for shorter time after that wakes up and start the same procedure from beginning. If successfully polled then it collects data from all sensors, convert this data into a packet

format and send it to the gateway node along with its node ID and checksum, refer to Table. II for packet structure. Gateway node after receiving packet will check for node ID in its database and will perform checksum match. In case of any error in the packet, it will be discarded and gateway node will move back to polling mode. For overcoming local communication issue re-transmission of packet at sensor node has been enabled. In case, if local communication does not take place successfully then sensor node can re-transmit the packet up to certain limit as defined by authors. If timer expires then the sensor and gateway node will automatically be re-started. If gateway node successfully received the packet without any errors in it then it will send an acknowledgment packet back to sensor node. The sensor node after receiving acknowledgment will move back to sleeping mode and gateway node will start making packet for GPRS module. A timer for GPRS module has also been defined because sometimes connection does not get established. If successfully established then the data collected from sensor node will be transferred to ThingSpeak cloud. Please refer to Fig. 3 & Fig. 2 for in depth understanding of complete flow that is enabled at sensor and gateway nodes. It must be clear that authors have tried their best to implement reliability and optimal performance under different scenarios so that either the sensor or gateway node does not get stuck in any long/ never ending loops.

Fig. 6. Testing of nodes at roof of CIS department NED UET, Pakistan

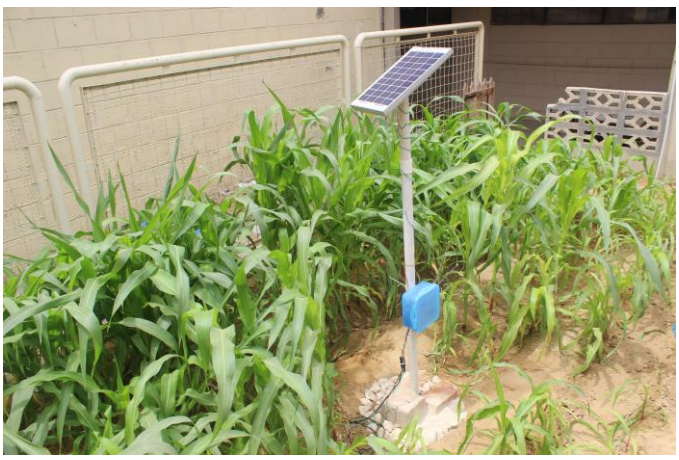


Fig. 7. Testbed deployed at NED UET Karachi, Pakistan

IV. TESTBEDS DEPLOYMENT & RESULTS

Before casing, nodes were tested in laboratory over month duration as shown in Fig. 5. All the minor bugs observed in software during this test were resolved. After properly casing the nodes were again tested by deploying at the roof of Computer & Information Systems Engineering (CIS) department of NED University over few days as shown in Fig. 6. After ensuring reliable operation of designed nodes under controlled environments, two sensor and a gateway node, as described in section II-B1, were deployed at a small farm of corn crop in Civil department of NED University during the month of June, 2016, please refer to Fig. 7. Though during this time period the weather of Karachi was very cloudy but the deployed nodes performed very well and authors were able to observe real time readings through ThingSpeak cloud, refer to Fig. 9. During the month of August, 2016 the testbeds with modified gateway node, as mentioned in section II-B2, were deployed in a bitter gourd farm at Gadap, Karachi, please refer to Fig. 8.

Fig. 8. Testbed deployed at Gadap Karachi, Pakistan

The testbeds have generated numerous amounts of data since their deployment which will be used in future research for introducing automation in agricultural sector by exploiting machine learning techniques.

Fig. 9. Real time readings received at cloud through testbed deployed at civil department NED UET

V. Conclusion

Although results obtained in real time are following certain pattern throughout the day. However, introducing automation in agricultural sector is challenging because of the environmental conditions which are continuously changing and beyond human's control. Through testing under different scenarios and in depth analysis, the performance of nodes was optimized. Therefore, their performance was satisfactory when deployed in open fields which were posed to continuously varying environments. Future research includes exploitation of obtained database of monitored parameters through machine learning. After that automation will be introduced in agricultural sector for efficient utilization of water resource.

Acknowledgment

This project is funded by DAAD, Germany under Pakistan German research program. It has been carried out in collaboration with DFKI, German Research Centre for Artificial Intelligence. The authors would like to thank Prof. Dr. Muhammad Shafqat Ejaz for cooperation during deployment of nodes at civil department NED UET.

References

[1] (2015, Oct) The internet of things: An overview, understanding the issues and challenges of a more connected world. Internet Society. [Online]. Available: https://www.internetsociety.org/sites/default/files/ISOC-IoT-Overview-20151014 0.pdf

[2] (2015) Cloud and mobile network traffic forecast - visual networking index (vni). Cisco. [Online]. Available: http://cisco.com/c/en/us/solutions/serviceprovider/visual-networking-index-vni/index.html

[3] "Global connectivity index," Huawei Technologies Co., Ltd.,, Sept 2015. [Online]. Available: http://www.huawei.com/minisite/gci/en/

[4] Manyika, James, M. Chui, P. Bisson, J. Woetzel, R. Dobbs, J. Bughin, and D. Aharon, "Global connectivity index," McKinsey Global Institute, June 2015.

[5] T. Kalaivani, A. Allirani, and P. Priya, "A survey on zigbee based wireless sensor networks in agriculture," in *3rd International Conference on Trendz in Information Sciences Computing (TISC2011)*, Dec 2011, pp. 85–89.

[6] J. Ma, X. Zhou, S. Li, and Z. Li, "Connecting agriculture to the internet of things through sensor networks," in *Internet of Things (iThings/CPSCom), 2011 International Conference on and 4th International Conference on Cyber, Physical and Social Computing*, Oct 2011, pp. 184–187.

[7] P. Asolkar and U. Bhadade, "Analyzing and predicting the green house parameters of crops," *International Journal of Computer Applications*, vol. 95, no. 15, 2014.

[8] S. P. Singh and S. Sharma, "A survey on research issues in wireless sensor networks," *OPEN TRANSACTIONS ON WIRELESS SENSOR NETWORK*, vol. 2, no. 1, pp. 1–18, 2015.

[9] S. Li, J. Cui, and Z. Li, "Wireless sensor network for precise agriculture monitoring," in *Intelligent Computation Technology and Automation (ICICTA), 2011 International Conference on*, vol. 1. IEEE, 2011, pp. 307–310.

Estimating News Coverage Index Using Probabilistic Topic Model

Muazzam Ahmed Siddiqui[1], Tafseer Ahmed[2], Sohail Abdul Sattar[3]

[1] Department of Information Systems, Faculty of Computing & Information Technology
King Abdulaziz University Jeddah, Saudi Arabia Email maasiddiqui@kau.edu.sa
[2] DHA Suffa University Karachi, Pakistan, Email tafseer@dsu.edu.pk
[3] NED University of Engineering & Technology Karachi, Pakistan, Email sattar@neduet.edu.pk

***Abstract*—We present a system to estimate the news coverage index using probabilistic topic model. The news coverage index reflects the coverage given to a specific topic by a newspaper. A topic is defined as an unordered and unlabeled collection of words that are more likely to be found together in a given context. A probabilistic topic model is a statistical algorithm that extracts these topics from a collection of documents. The coverage index for a topic was computed as a function of the number of such documents carrying that topic. As a proof of concept, we developed a prototype to estimate the coverage given by the editorial and op-ed pieces in two Urdu newspapers, Nawaiwaqt and the Daily Express. The index was analyzed using a multidimensional data model, which enabled us to view the coverage from different dimensions. Visualizations are provided in the form of word clouds and a author-topic graph to compare the coverage index of the two newspapers. The index can be used to determine what is covered or not covered and identify the agenda as well as specific biases associated with newspapers and writers.**

Keywords—news coverage index; probabilistic topic model; multidimensional data model; computational journalism; urdu natural language processing

I. INTRODUCTION

The Project of Excellence in Journalism (PEJ) developed the News Coverage Index to quantify the news coverage by the US media [1]. This, in turn, determines the broad news agenda of different media outlets.

Topic modeling discovers topics from a collection of documents. A topic model, also referred to as probabilistic topic model is an algorithm that discovers the hidden thematic structure from a collection of documents. The discovered topics differ from the traditional sense of the term as a topic extracted from a topic model is an ordered collection of words, which are more likely to be found together within a given context. The main idea behind topic modeling is to consider a document as a distribution over topics and a topic as a distribution over words found in the document collection. Therefore, each document may contain multiple topics with different probabilities and each topic may span multiple documents. Similarly, each topic contains multiple words with different probabilities and each word may occur in different topics. The end product is a set of collections of coherent words, each representing a topic and a topic assignment to each document representing the most probable topic for that document. The word collections are unlabeled as the algorithm only assigns an ID to each collection.

This paper proposes an application of a probabilistic topic model to estimate the news coverage index of selected Urdu newspaper. The algorithm was used to extract topics from selected Urdu newspapers and the index was estimated as a function of the number of documents discussing each topic. A multidimensional data model was used to analyze the index.

Urdu is the national language of Pakistan, and is also used to read and write in many parts of India and other South Eastern regions. It is written in Arabic script and is inherently cursive in nature. Text Mining on Urdu text has only recently attracted the researchers' interest and not much published work is available on the subject. To the best of our knowledge, we were not able to find any published work applying topic modeling to Urdu newspapers.

To perform topic modeling on Urdu text, more than twelve thousand documents consisting of editorials and columns from the repositories of two popular Urdu daily newspapers, Express [2] and Nawaiwaqt [3], were automatically crawled and cleaned to make the corpus. Each document in the corpus was indexed with the name of newspaper, author, the date of publication and source URL.

In topic modeling we tried three different granularities of topics i.e. we extracted 10, 25 and 50 topics from our editorial corpus. We perform different kinds of analysis e.g. topic vs. newspaper, topic vs. author and topic vs. month vs. newspaper.

II. PROBABILISTIC TOPIC MODEL

Blei and Lafferty [4] proposed the development of tools that can make web browsing and the use of digital libraries more efficient. These tools would use the methodologies adapted for topic modeling that extract meaningful structures from otherwise unstructured data sets.

It is common experience that most documents contain content that has a variety of relevant topics and underlying hidden concepts that may not be explicitly obvious. The proposed scheme suggests the extraction of multiple topics from such documents that may look unstructured but has the potential to provide the researcher a number of sub topics and relevant ideas that are interconnected in varying degrees with the main theme. A topic model that is merely able to determine a single theme or topic would thus not be considered efficient compared to those that is able to derive highest possible number of topics and sub topics.

Thus a successful model would identify and extract all the previously unidentified and unfamiliar topics from the document creating a new piece of information from a present set of unstructured data. This is achieved by the introduction of a hidden variable model used by the LDA, The Latent Dirichlet Allocation, in the data under consideration. Finally a posterior probabilistic inference is used to draw out the hidden patterns and structures.

The hidden variables are distributed as latent topical structures, the topics that are sought out while the LDA decomposes the text on which these models are to be applied. All the words the document is composed of are the data observed while hidden variables constitute the topics that are to be found out or discovered. The interaction of the models applied to the unstructured data produces information through a probabilistic generative process, called topic modeling.

The structures and information thus derived enable better digital browsing and more efficient data retrieval.

David Blei et. al. [5] presents the algorithm and mathematical details of LDA. The general concept governing LDA is that all documents necessarily are composed of words, and all words refer to latent topics and relate to the theme in many ways. Therefore the main idea under consideration here is text processing though the LDA may be used for many other purposes.

Error! Reference source not found. presents the LDA in plate notation. We may observe the three hierarchical levels where the outer one presents examples of documents while the topic words in the document are presented as the inner plates. The globally fixed parameters α and β are required to be sampled only during corpus generation.

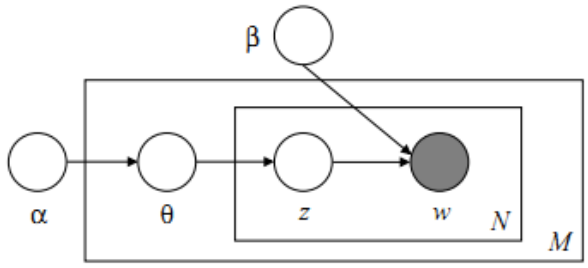


Fig. 1. LDA represented using the plate notation

α = arbitrary parameter for Dirichlet prior for every document distribution

β = arbitrary parameter for Dirichlet prior for every topic word distribution

θ_i = random variable for topic distribution in i^{th} document.

z_{ij} = variable for topic of the j^{th} word in the i^{th} document.

w_{ij} = the only observed variable - a specific word in the i^{th} document.

A word w^i is an item drawn from vocabulary indexed with superscript i. A document is treated a vector of N words $\boldsymbol{w} = (w_1, w_2, w_3, \dots, w_N)$ where w_n is the nth (subscript) word in the ordered tuple. A corpus is a set of M documents $D = \{w_i, w_i, w_i, \dots, w_N\}$

LDA performs the following steps for each $\boldsymbol{w}$ in D:

1. Sample N from Poisson(ξ)
2. Sample θ from Dir(α)
3. For each of the N words w_n
    a. Sample topic z_n from Multinomial(θ)
    b. Sample word w_n from $p(w_n|z_n, \beta)$

The probability of corpus is given by:

$$p(D|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d)\, p(w_{dn}|z_{dn},\beta) \right) d\theta_d$$

III. DATA

Data for this work was collected from two Urdu newspapers, the Daily Express and the Nawaiwaqt. In the beginning, we explored several Urdu newspaper websites to find columns and editorials. Unfortunately,

most of these websites offer their content in the form of images instead of text[1]. To analyze the text an Optical Character Recognition (OCR) software is required. To the best of our knowledge, no Urdu OCR tool is publicly available and building an OCR was beyond the scope of our work. As a result we limited our attention to the newspapers that offer their contents in text format. We were only able to find four Urdu newspaper websites. These include the Daily Express, the Nawaiwaqt, BBC Urdu and the Jang. Among these four newspaper websites, Jang's editorial and columns section was inaccessible at the time of data collection. For BBC Urdu we were not able to locate the archives.

To collect the data from the Daily Express and the Nawaiwaqt, we used a crawler to download the contents. The following steps were involved in the process

1. Explore the site structure manually to identify the URLs containing columns and editorials
2. Use a crawler to fetch the web pages from the aforementioned newspaper websites
3. Clean the HTML pages and extract the text of the column and/or editorial
4. Clean up the text collection of unwanted documents that may have been added due to inconsistent or improper website structure

A. Corpus

The original corpus consisted of 27302 documents including columns, editorials, news items, special reports and pages containing Iqbal's[2] poetry from December 2008 to May 2014. Each of these documents was stored as a separate UTF-8 encoded text file. In addition, each document was indexed with the author, title, date/time and the source URL. The index was stored in a separate file. Since the focus of our work was on columns and editorials only, we manually separated the required documents from the rest. Our final corpus consisted of 12058 documents containing columns and editorials.

B. Preprocessing

Preprocessing of corpus included tokenization, removal of punctuation and any leftover HTML elements and removal of Urdu stopwords. The preprocessing was done in Java. A list of Urdu stopwords was prepared by identifying the most frequent function words. In Urdu these include اس {this}, اور {and}, پھر {then}, تھا {was} etc. **Error! Reference source not found.** displays the corpus statistics after preprocessing. The number of documents from the Express is more than twice of the Nawaiwaqt. Any further analysis must take this fact into account. The number of tokens (words) displays similar pattern. On the other hand, the number of types (distinct words) slightly deviate from this pattern and the type/token ratio is higher in the Nawaiwaqt than in the Express. A quick analysis revealed that the number of contributing columnists for the Nawaiwaqt is much more than the Express, resulting in a more diverse use of words and thus a higher type/token ratio. The average document size is almost similar indicating a similar length of column or editorial used by both the newspapers.

Statistics	Express	Nawaiwaqt
Documents	8,208	3,850
Tokens	4,593,539	2,166,566
Types	128,953	82,165
Type/Token Ratio	0.028	0.038
Avg Doc Size (tokens)	559.6	562.7

TABLE I. CORPUS STATISTICS

C. Vector Space Model

The probabilistic topic modeling algorithm, LDA, used to build the topic models, require the input data to be in a document term matrix format. A document term matrix is an implementation of the vector space model where each document is represented as a vector in an n-dimensional space, where n is the total number of terms in the corpus.

The vector space model is a bag of words model where the position of the term in a document is ignored. The document term matrix carries the documents in rows and terms in columns. The intersection of each row and column contains a number representing the frequency or presence of the term in the document. While there are other representations also available, the topic modeling algorithm requires the term frequency representation. To create the document term matrix, we used the *tm* package [6] in R [7]. For tokenization we used the default Arabic tokenizer in the tm package, as there is no Urdu tokenizer available. The document term matrix was created from the entire corpus containing the documents from both the newspapers. Because of the tokenizer difference the number of tokens and types is different.

A comparison of the tokens obtained previously using Java and later the *tm* package in *R* revealed the difference in mostly less frequent terms. The tokenizers were mostly consistent otherwise. It is important to mention at this step that the topic models were created from the entire corpus, instead of each newspaper separately. The presence of the document index made it possible to trace the document to the author, title, newspaper and time for the analysis later.

[1] The text is originally written using an Urdu wordprocessing software but is provided as an image on the newspaper website to circumvent font and encoding problems at the client side

[2] Mohammed Iqbal is the national poet of Pakistan and his couplets are featured regularly on different newspapers

D. Data Reduction

Data reduction was applied to obtain a smaller representation of the data. The main purpose of data reduction is to remove noise terms and select the most important terms. Even though the tokenizer available in the tm package resulted in 2.9 million tokens compared to the 6.7 million tokens from the Java tokenizer, the number of dimensions was still prohibitively large resulting in a sparse document term matrix. We removed the sparse terms from the matrix by setting up a maximum allowed sparsity threshold of 0.99 and removed the terms with a sparsity value of more than the threshold. This greatly reduced the number of dimensions but also decreased the sample size to 6029 documents as a number of documents consisted of sparse terms only.

Statistics	Number
Documents	6029
Tokens	2,052,362
Types	2,927
Type/Token Ratio	0.001
Avg. Doc Size (tokens)	340.4

TABLE II. CORPUS STATISTICS AFTER PREPROCESSING

IV. NEWS COVERAGE INDEX

We used the *topicmodels* (B Gruen and K Hornik, 2011) package in R to extract the topics from the corpus. For each of the extracted topics, we computed the news coverage index.

A. Index Computation

The news coverage index for a topic was computed as a normalized count of documents belonging to that topic.

$$I_t = \frac{C_t}{N}$$

Where I_t is the index for topic t, C_t is the number of documents containing t as the major topic and N is the total number of documents.

B. Multidimensional Data Model

A multidimensional data model enables the data to be viewed from multiple dimensions or perspectives [(Han, Kamber, & Pei, 2011)]. Each of these dimensions supports a hierarchy enabling data summarization at different levels of granularities. We used the entity and the time dimension to view the index. The entity dimension has three levels in the hierarchy (all -> newspaper -> author), while the time has four levels (all -> year -> month -> day). It should be noted that the data model was not implemented as a data cube in a data warehouse. The OLAP operations were implemented in Microsoft Excel using Pivot tables.

V. RESULTS AND DISCUSSION

Once computed in Excel, we used the *wordcloud* [9] and *igraph* [10] packages in R for index visualization, in addition to the charts provided by Excel.

The LDA algorithm used in the *topicmodels* package requires the number of topics as an input parameter. This number defines the granularity level of the topics. Smaller number gives coarse grained topics while a larger number results in fine grained topics. To explore this we extracted three sets of topics. The sizes of topic sets are 10, 25 and 50 topics. The most important (probable) topic of each document is also extracted. Following is a brief description of the results of further analysis of topics and document features (newspaper, author, date). All of the following examples are samples from a big list of results, as we do not present all the results due to lack of space.

We also present some discussion about the interpretation of the resulting tables and graphs. It is important to note that in contrast to objective data, the interpretation and discussion are subjective and depends on authors' or one of the author's subjective assessment. The social scientists, current affair analyst and other experts can explore much better interpretations from this data.

A. Topic Coverage by Newspaper

Error! Reference source not found. displays the top four words, that we would consider the representative of each topic when the algorithm was run to extract 10 topics from the corpus.

Id	Top 4 words
1	کام, فلم, بات, کہا *kAm* (work), *film* , *bAt* (matter), *kahA* (said)
2	سیاسی, پارٹی, عوام, انتخابات *siyAsI* (political), *party*, *awAm* (people), *intikhabAt* (election)
3	روپے, ٹیکس, پاکستان, ایف *rupay* (rupees), *tax*, *pakistan*, *F*
4	حکومت, شریف, وزیر, نواز *hukomat* (government), *sharif*, *vazIr* (minster), *nawaz*
5	کراچی, پولیس, سندھ, افراد *karachi*, *police*, *sindh, afrAd* (people)
6	پاکستان,بھارت,بھارتی,کشمیر *Pakistan*, *bHArat* (India), *bHArtI* (Indian), *kashmir*
7	پاکستان, دہشت, طالبان, حکومت *Pakistan*, *dahshat* (terror), *taliban*, *hukomat* (government)
8	ہمارے, ملک, لوگوں, صاحب *hamaray* (our), *mulk* (country), *logon* (people), *sAhab* (Mr.)
9	جنرل, مشرف, صدر, پرویز *general, Musharraf*, *sadar* (president), *Pervaiz*
10	ملک, دنیا, ترقی, نظام *mulk* (country), *dunyA* (world), *taraqqi* (development), *nizAm* (system)

TABLE III. TOP FOUR WORDS IN EACH TOPIC WHEN 10 TOPICS WERE EXTRACTED

Majority of these set of words are easy to interpret for an Urdu newspaper reader or a person interested in Pakistan's current affairs.

The increase in number of topics in other experiments gives two effects. Some new topics are added which are not covered in the above list. Some topics appears that consist of left-over stop words because of omission, tokenization or normalization problem.

We try of find whether both newspaper give similar coverage to all topics or one newspaper has more or less focus on a topic. **Error! Reference source not found.**, **Error! Reference source not found.** and **Error! Reference source not found.** display the topic coverage comparison using the news coverage index for 10, 20 and 50 topics respectively. The topics returned from the LDA algorithm were unlabeled; we concatenated the first four terms in the topic to assign a label to each topic for an easier interpretation.

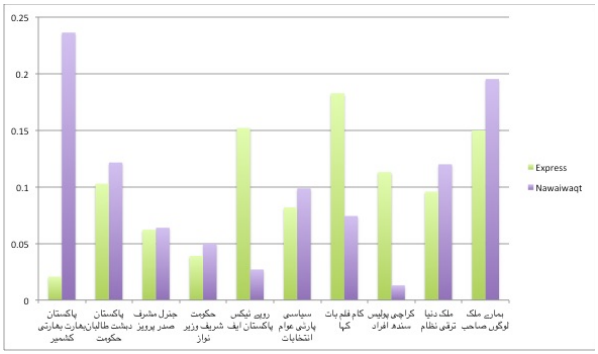


Fig. 2. News coverage comparison for 10 topics using the index

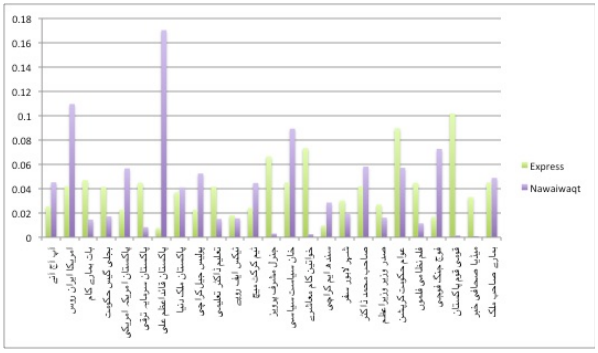


Fig. 3. News coverage comparison for 20 topics using the index

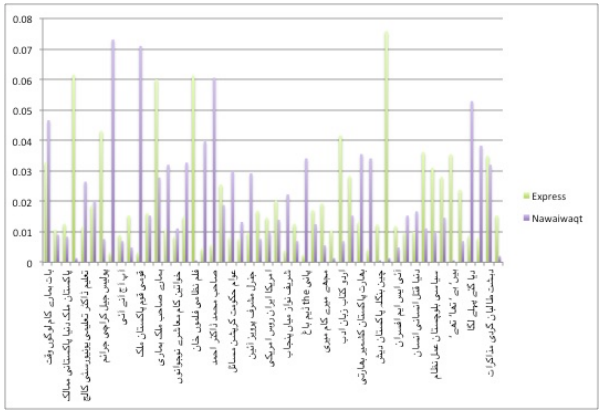


Fig. 4. News coverage comparison for 50 topics using the index

The figures show that the coverage varies from topic to topic. Some topics receive similar coverage from both the newspapers, while some topics were mainly covered by only one of the two newspapers. The same topic coverage is shown as a word cloud comparison in **Error! Reference source not found.**. The size of the font reflects the frequency of the topic in the newspaper. The word cloud is displayed only for the 10 topics case.

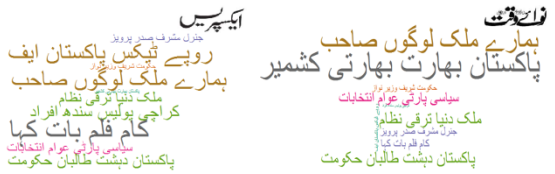


Fig. 5. Word clouds comparing the topics from the Express and the Nawaiwaqt

B. *Topic Coverage by Authors*

We drilled down into our data model on the entity dimension to determine the topic coverage per author. The results are provided in the form of an author topic graph (see Appendix A). In the graph, the authors and topics are represented as nodes. An edge between an author and the topic node indicates that the author discussed the topic in his/her columns. The width of the edge represents the coverage of the topic by the author. The authors that have fewer documents or the edges that have small weights are removed to simplify the visualized graph. The author topic graph was created for the topic model generated for 50 topics. The topics are represented by their top three words enclosed in curly brackets.

The authors occurring closer (usually) write about similar topics. The distance between the topic and the author shows the strength (frequency) of the author's relation with that topic. The author who write on similar topics are clustered around that topic.

We find some observation about the relation of author to the topics. We focus, as a sample, on three authors Intizar Hussain, Javed Chaudry and Talat Hussain.

Intizar Hussain was fundamentally a fiction writer. In 10 topic analysis, 67.8% of his columns belonged to topic 1 and 32.2% belonged to topic 8 (Consult table 5.1 for most frequent words of these topics). Both of these topics are un-interpretable (at least to the authors of this paper) and can be labeled as miscellaneous topic. In 50 topic analysis, 66.7% of his columns are related to topic 33 whose most important words are: *Urdu*, *kitAb* (book), *zabAn* (language), *adab* (literature). Hence, by increasing the number of topics to 50, we discovered a topic related to literature.

Javed Chaudry writes in his unique style mostly about social issue. In 10 word and 50 word analysis, he is related to the topics which can be labeled as miscellaneous. In 50 topic analysis, we find a miscellaneous topic (id 45) that seems to be related only to his articles. 88% of Javed Chaurdry's articles are related to topic 45. Only one more author's colums are related to topic 45, and his percentage is 13%.

Talat Hussain is a political analyst. The topic analysis gives the following statistics: topic-7 (28.6%), topic-8 (25.4%), topic-2 (12.7%), topic-10 (9.5%) and topic-9 (9.5%). The most important words of these topics can be consulted from table 5.1. The statistics for 50 topic analysis are: topic -1 (14.3%), topic-49 (12.7%), topic-23 (11.1%), topic-9 (9.5%) and topic-27 (6.3%). The numbers show the variety of topics by this author.

C. Topic Trends

The second dimension used to compare the two newspapers was time. We summarized the news coverage index at the month level and identified a time period common for both newspapers. The oldest document from the Nawaiwaqt was from December 2008 while the oldest document from the Daily Express dated back to July 2012 only. For the comparison purpose, we chose the time span from July 2012 to May 2014. **Error! Reference source not found.** and **Error! Reference source not found.** display the topic trends from July 2007 to May 2014 for two different topics for the two newspapers. The topic related to General Parvez Musharraf shows a somewhat similar trend by the two newspapers' columns till 2013 in **Error! Reference source not found.**. From the beginning of 2014, the topic was given a lot of coverage by the Nawaiwaqt. On the other hand, **Error! Reference source not found.** displays a very different coverage to the topic of law and order situation in Karachi by the two newspaper columns.

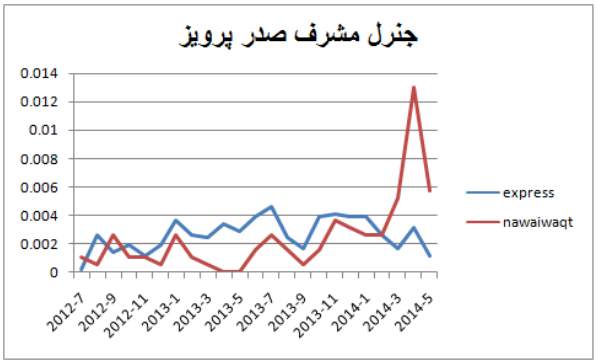


Fig. 6. Comparison of the coverage of the topic related to General Parvez Musharraf

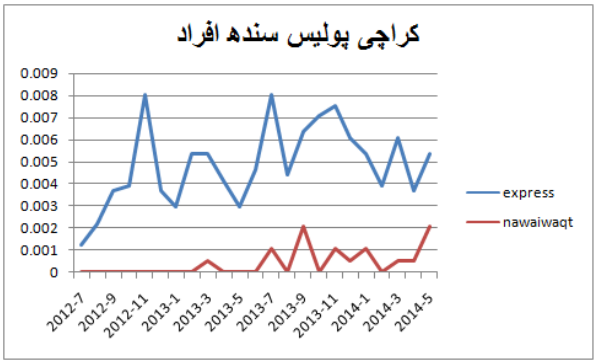


Fig. 7. Comparison of the coverage of the topic related to the law and order situation in Karachi

D. N-gram analysis

The topic models were created from the unigrams, i.e. single word tokens. This resulted in breaking up of a number of multiword expressions, especially named entities, into single tokens. A solution to this problem is to identify named entities first and treat them as a single term. Building a named entity recognition system to mark the named entities and performing entity resolution was beyond the scope of this work. However, we identified collocations through a bigram and trigram analysis of the corpus. We could not integrate this component in our topic analysis (due to lack of time). However, the bigram and trigram clouds showed promising (multiword identification) results. **Error! Reference source not found.** displays the word cloud comparison of trigrams extracted from each newspaper separately.

Fig. 8. A frequency comparison of the trigrams (three word combinations) of the Daily Express and the Nawaiwaqt

VI. CONCLUSION

A prototype to determine the topic coverage by newspapers was presented in the paper. The prototype consists of two main components. The first component estimates a news coverage index, a quantitative value, computed through a probabilistic topic model and identifying how much coverage is given to any specific topic. The second component uses a multidimensional data model to analyze the index across different dimensions. In this work we used only two dimensions and implemented the data model using Microsoft Excel's pivot tables. This proof-of-concept implementation can be replaced by building a full-fledged data warehouse with additional dimensions and a suite of OLAP operations. In this work, the index was computed and analyzed to compare the topic coverage by two Urdu newspapers from Pakistan. Other languages can be used by replacing the tokenizer and additional language dependent components such as named entity and/or multiword expression identifiers, if used.

VII. REFERENCES

[1] "News Coverage Index Methodology," Pew Research Center, 25 8 2011. [Online]. Available: http://www.journalism.org/news_index_methodology/99/. [Accessed 24 8 2016].

[2] [Online]. Available: http://www.express.pk/.

[3] [Online]. Available: http://www.nawaiwaqt.com.pk/.

[4] D. Blei, J. Lafferty., "Topic Models," *Chapman & Hall/CRC Data Mining and Knowledge Discovery Series,* 2009.

[5] Andrew Y. Ng, Michael I. Jordan David M. Blei, "Latent Dirichlet Allocation," *Journal of Machine Learning Research 3,* pp. 993-1022, 2003.

[6] I Feinerer, K Hornik, and D Meyer, "Text Mining Infrastructure in R," *Journal of Statistical Software,* vol. 25, no. 5, pp. 1-54, 2009..

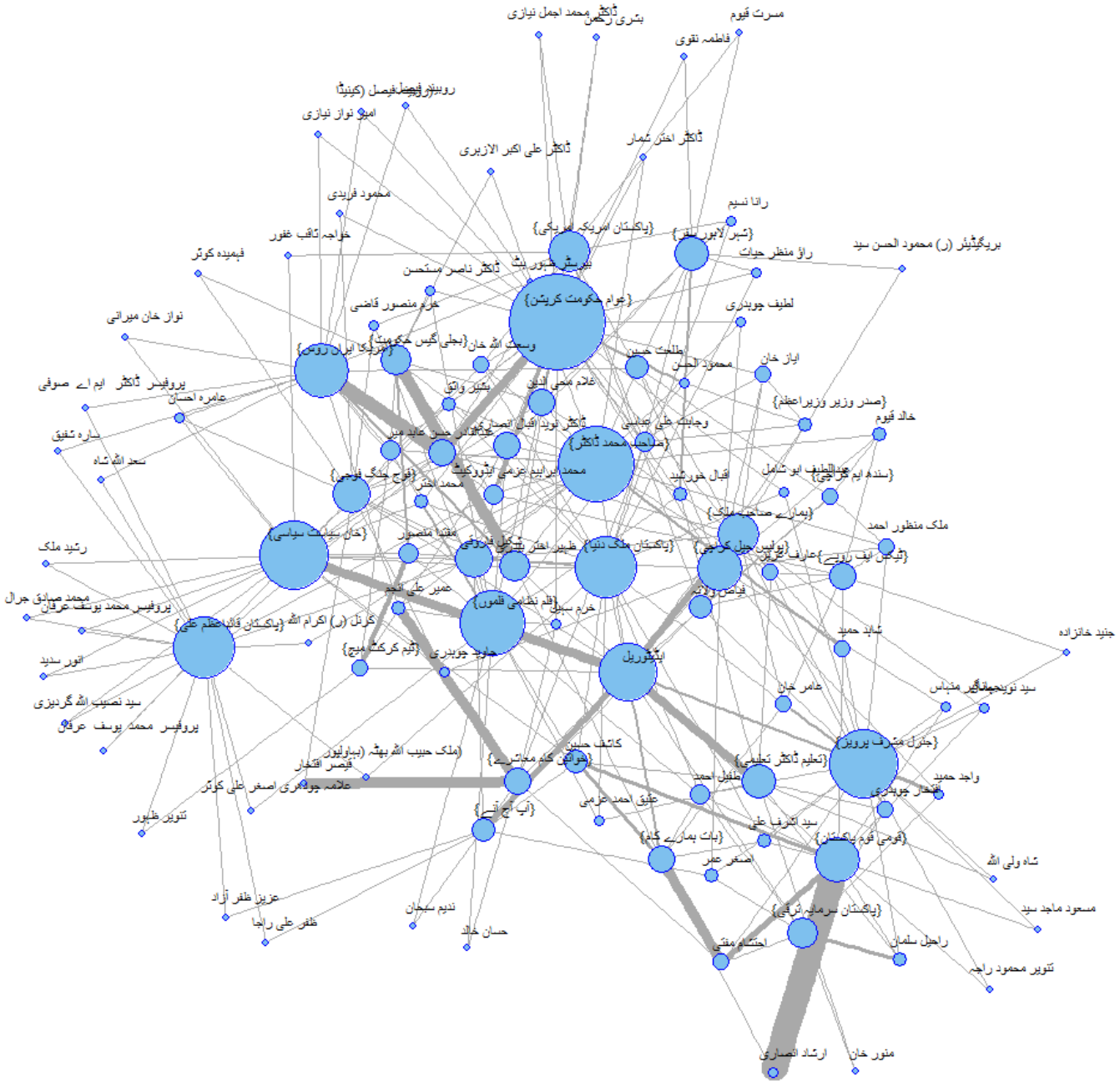
[7] R_Core_Team, *R: A Language and Environment for Statistical Computing,* Vienna: R Foundation for Statistical Computing, 2013.

[8] B Gruen and K Hornik, "topicmodels: An R Package for Fitting Topic Models," *Journal of Statistical Software,* vol. 40, no. 13, 2011.

[9] I. Fellows. [Online]. Available: http://cran.r-project.org/web/packages/wordcloud/wordcloud.pdf.

[10] G. Csardi and T. Nepusz, "The igraph software package for complex network research," *InterJournal,* vol. Complex Systems, p. 1695, 2006.

APPENDIX A: AUTHOR TOPIC GRAPH



ICONICS 2016 POSTERS

Does Social Media Impacts Learning? An Empirical Study.

Hidayatullah Shaikh
Department of Computer Science Shah Abdul Latif University, Khairpur, Pakistan

Wazir Ali Jamro
Department of Computer Science Shah Abdul Latif University, Khairpur, Pakistan

Javed Ahmed Mahar
Department of Computer Science Shah Abdul Latif University, Khairpur, Pakistan

Arif Hussain Magsi
I.T Centre, Sindhi Agriculture University, Tando Jam, Sindh, Pakistan.

Altaf Hussain Bhatti
Physics Department Shah Abdul Latif University, Khairpur, Pakistan.

***Abstract*— Social Media has become integral part of our daily lives but still most of the societies are unknown to take its educational advantages. Therefore, there is need to explore educational insights, usage priorities and impact of these Social Networking Sites (SNSs) at higher education level because a massive number of youth is already engaged on social media platforms. In this regard, present study aims to assess the impact of Facebook usage on learning strategies at university level. Because it is most influenced social site across world including Pakistan. Unfortunately, most of the people and students consider that using the SNS is a wastage of time even at higher education level. Therefore, we have reviewed the literature in order explore educational charms of SNSs and their global adaptation in teaching and learning practices. Moreover, we have conducted survey in order to investigate the real usage priority of Facebook among undergraduate students of 5 different universities. In the end survey data is analyzed and results have been presented regarding students' priorities. This work will benefit the educationists and students in order to adopt emerging e-learning trends.**

Index Terms— Web 2.o, Social-Media, Facebook, E-learning, Survey, and Higher Education.

I. INTRODUCTION

Internet has become very sophisticated platform for anyone to share information, create and upload with the Web 2.0 technologies, therefore todays communities are more connected than ever before. Web 2.0 technologies are becoming popular and embedded in daily life routines. Particularly, among young generation in many places and other different social settings. The term includes Social Networking Sites (SNS), video sharing, blogs, wikis etc. Social media platforms offer marvelous learning facilities in every stage of life. Learning trends are changing rapidly, therefore technological methods for effective teaching and learning are dominating old dated traditional methods. Hence, Social Media is one of the emerging technology which has been largely integrated, experienced and implemented globally in higher education. The social media applications bear a lot of educational insights. In this regard many studies have been carried out in order to investigate the impact and usage priorities of social media including Facebook for the learning enhancement. These studies consist on some specific university students, certain group of students, experiments and online surveys etc. Through these studies it is learnt that social media users not only access the rapid and large amount of data but also create content (text, writing, photo and video), react and remain involved individually/ collaboratively for learning purpose on the social networking sites. Precisely Facebook provide a blended virtual and real learning environment specially at higher education level. The important educational features provided by the SNSs including Facebook are; connectivity, information discovery and sharing, content creation and modification and knowledge/information aggregation. Therefore, we have presented literature review and conducted survey of 5 different universities of Sindh province, in order to analyze the usage priorities of university students. The Facebook is selected for this study because, it is most popular SNS across the globe. Currently more than 800 million people have joined Facebook in which 12.5% are active users in Pakistan. Therefore, this study is carried out in order to investigate authentic use of Facebook among university students. This paper is organized as follows; in section 2 we have presented literature review of SNSs including Facebook, section 3 elaborates the adopted research methodology and analysis of survey data is presented in 4th section. The 5th consists of discussion and some suggestions regarding positive use of Facebook and finally paper is concluded in 6th section.

II. LITERATURE REVIEW

A Personal Learning Environment (PLE) is a promising potential approach for the integration of formal and informal learning by using social media platforms and supporting self-regulated learning for students in higher education (Dabbagh and Kitsantas,.2011). Facebook as a learning platform, it

includes constructivism which invokes learning environments that are community-centered, Knowledge-centered, Learner-centered and Assessment-centered. Learning is a social activity resulting from a combination of interactions, collaborative activities and communication with others (Vygotsky; 1978). Which is believed to enhance critical thinking and cognitive skills. Facebook provide online learning platform that involves learning community with certain learning goals and outcomes to accomplish. It has been found evidently that the exchange of ideas and knowledge among participants in a learning community increase the motivation to learn (Hew; Koh; and Herring; 2010) and (LaRue; 2012). Social-media has influence the lives from home, school and workplace (A. Nawaz, G. Abbas, M. F. Javed, N. Mughal and M. Nabeel., 2015). Colleges and universities are embracing social media platforms and realizing the potential and implications of its usage (Rachel Reuben). According to (Mark Blankenship., 2010), the major inter-connected literacies of SNSs are Attention, participation, collaboration, network awareness and critical consumption. Facebook has numerous applications that support teaching and learning. It has been found to cultivate positive learning experiences as well as enhance connection between teacher and student (Mazer et al., 2007). The SNS allow for better collaboration between students, in line with the educational standard of social- constructivism (Huijser; 2008). Precisely, this collaborative potential of SNS technologies (Kelm; 2011) has led to calls for their adoption in teaching and learning. It is also found by (Baran; 2010) in a study of Turkish university students, that "more than 60% of students are communicating with their classmates through a module Facebook group helped, which stimulate them in their learning. However, the distinction is drawn between tutor-initiated and student-initiated uses of Facebook, with the most successful uses being those created by students themselves (Whedon; 2011), while tutor-initiated uses were seen as somewhat fake (Ipsos MORI; 2008). It appears that students see some benefit in using SNSs such as Facebook for certain aspects of peer communication, often focused around group work (Selwyn; Madge et al., 2009). The keys to effective learning in social-networking environments have been identified by (Salavuo; 2008) as participation, presence and ownership, with (Conole et al.2008) adding that students should have a sense of self control. According to (Mullen and Wedwick; 2008), the previous definition of literate was defined as a person who being able to read and write, but the literate of this century must be able to download, upload, rip, burn, chat, save, blog, Skype, IM, and share. A study on the learning of language among university students in Thailand, revealed that usage of Facebook was indeed a beneficial secondary tool in enhancing the students' learning experience. It is claimed by (Ziegler; 2007) that SNSs have the ability to transform students from being passive learners to becoming active and intentional learners, which is indeed the principle of student-centered learning. Facebook has been suggested by (McCarthy; 2010) as an ideal host for a blended learning environment; as it was discovered to enhance peer relationships as students appreciated the interactive discussions that took place in the virtual-learning platform. The activity log students in their presented study showed an increase in learner engagement in the course, particularly with an assessment task. Recent evidence by (Ellison; Steinfield and Lampe., 2007; Kabilan, Ahmad and Abidin, 2010; Ng. and Wong., 2013) has shown that Facebook is one of the best way to enhance communication, instruct a more positive learning attitude, stimulate students to learn, encourage them to take their learning tasks more seriously and increasing their social capital via virtual interactions. The Facebook is practically observed as an ideal platform for informal learning by expanding course material in a group of various contemporary subjects which were not included in their class and to outside experts, thought leaders also included the option of participation with no interactive requirements. A way for learning strategy is also proposed by designing a professional communication model via social media (J. Cain, EdD, MS, and A. Policastri, PharmD, MBA 2011). The interviews conducted by (C. Pimmer, S. Linxen, Urs Gröhbiel) of medical students in various resource limited environments by considering the usage of Facebook in which some of the students appropriated these social media technologies as educational tool, with many users, students and medical professionals from across developing and emerging nations. The analysis of interviews, on clinical and medical topics revealed rich E-learning and E-teaching practices in informal-learning context. The social media has embedded a negative image upon many minds however it is also perceived as a pathway to develop vital knowledge and social skills among students beyond their campus walls (Wang., Chen and Liang,.2011). The social media has been widely adopted by students for personal interconnection. However, for teaching and learning purpose there is need to fully explore social media application (Chris Evans., 2014). A research conducted by (Hargittai, 2008a; Jones & Fox, 2009; Matney & Borland., 2009) shows that Facebook is most popular social site among college students, between 85 to 99% college students. An important study conducted by (Selami Aydin., 2012) suggest some practical recommendation regarding Facebook usage in educational environment. Moreover, six important aspects of using Facebook users, reasons for its usage, harmful effects, educational environment, effect on culture, language, education, subject variable and Facebook relationships are discussed. Facebook is very common and user-friendly technology that can improve the communication between teacher-students and student-student, which is not provided in the traditional classroom environment (McCharty., 2010). The educational service on the Facebook is known as face-learning (Ahmet Naci Coklar., 2012). The Academic engagement and peer interaction are two attractive features of Facebook as learning environment (McCharty., 2010), and it is an important learning platform, because of its popularity and ease of use. The views of the students taken by (Bosch., 2009) in a research study reveals that Facebook as an educational

environment is more useful for lesson or subject material sharing, project planning and it is an interactive platform between teachers and students. A study conducted by (Gray, K., Annabell, L., and Kennedy, G., 2010) shows that medical students use Facebook informally for learning enhancement in undergraduate lives. It is found by (Anam Ali 2016) in a study that Facebook is being used for assessment preparation, collaborative learning, providing support for creating peer connections, sharing educational resources and organizing studies, therefore, it is suggested that in some conditions Facebook can be adopted as a balancing educational platform that allows learners to create a personalized space for online learning among peers. Except this other various research studies have been carried out in which social media is probed for educational use, some of these practical research studies have been made by (Anderson., 2007; Madge et.al 2009; Bosch., 2009; Yuen & Yuen., 2008; Kosik., 2007; Solomon & Schrum., 2007).

II. Research Methodology.

The questionnaire was designed with multiple options which consists of 16 questions. The participants were n=1000 undergraduate students of 5 different universities including Iqra University Karachi, Federal Urdu University Karachi, Shah Abdul Latif Universities Khairpur, IBA Sukkur and Mehran University of Engineering, Science and Technology (MUET) Shaheed Zulifquar Ali Bhutto Campus, Khairpur. In a questionnaire the views of the students were taken in order to investigate the access medium, number of friends and followers, post sharing activities, access of subject related material, main purpose of using Facebook and their usage priorities, in order to enquire genuine use of Facebook among at higher education level. The number of male students is 73% and 27% female students participated in this survey respectively.

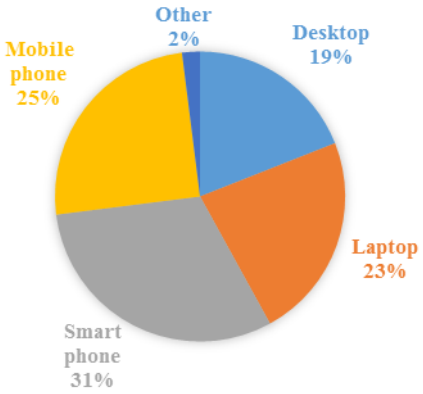
III. Analysis of Survey Data

For the analysis of data obtained through questionnaire, we counted the responses of survey participants and presented in percentage. All the multiple choices are added in graphs which demonstrate the activities or interests and usage priorities of students who participated in this survey.

a. *Access Mediums*

In this survey students were asked for their access medium and it is important to note that large number of students/ participants access the Facebook by using smart phone/ mobile phones. The figures 1. shows detail of access mediums of survey participants, in which it is noticed that mobile and smart phones are the most frequently used access mediums.

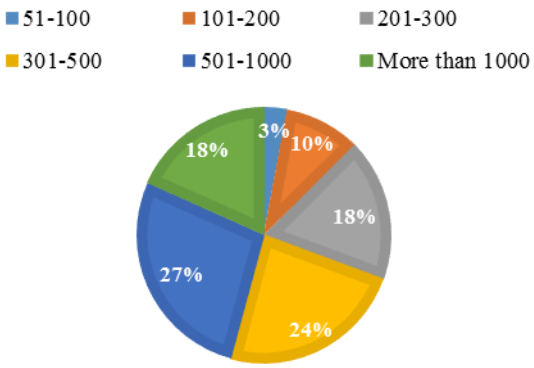
FIG I. GRAPH SHOWS FACEBOOK ACCESS MEDIUMS USED BY THE PARTICIPANTS.



b. *Connections (Friends/followers).*

The participants were asked in survey questionnaire that how much friends they have on the Facebook? In the response of this question it is revealed that a significant number of the participants have more than 500 and even more than 1000 friends, which is incredible number of connected people on a same platform. This number of connected people can collaboratively share and access the content/posts of each other that can equally benefit all the friends at one place in a uniform way. The detail of students' responses is presented in figure II.

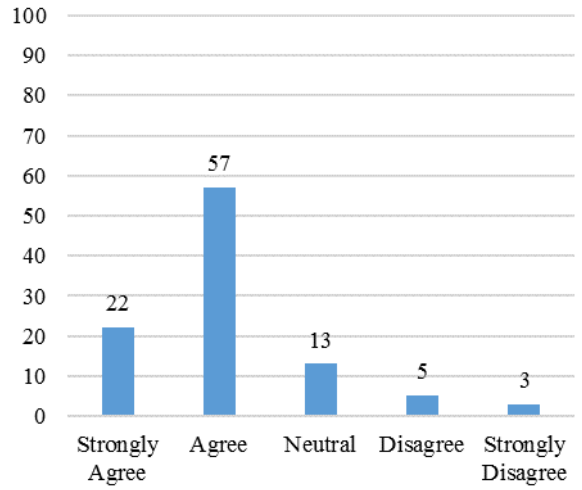
FIG. II. PRESENTS THE PERCENTAGE OF FRIENDS AND FOLLOWERS OF PARTICIPANTS.



c. *Facebook for content sharing.*

The participants/ students were asked that is Facebook an ideal tool for sharing contents? The significant number of participants 57% agreed upon the statements, the more details of students' responses is given in detail in figure 3.

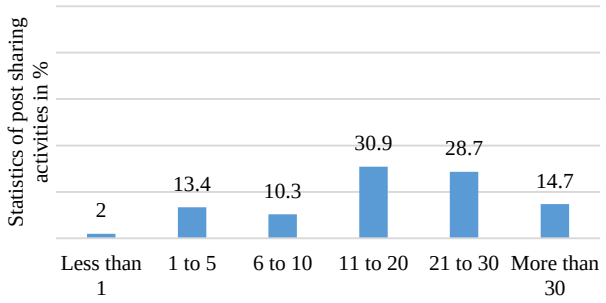
FIG. III. PERCENTAGE-WISE RESPONSES REGARDING FACEBOOK AS AN IDEAL TOOL FOR SHARING ACTIVITIES.



d. Activity log of content sharing

In the connection of above question we also asked about the Facebook usage in content or post sharing support. In figure no. 4, the responses of participants in graphical format are presented regarding access to the content or subject related material by using Facebook. This question was asked in order to record the weekly activity log of the survey participants which is presented in a graphical format.

FIG. IV. STATISTICS OF POST/CONTENT SHARING ACTIVITIES.

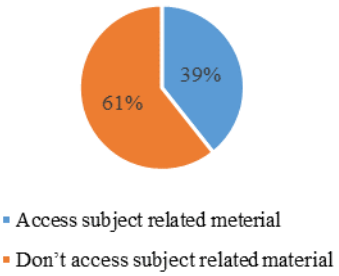


The figure 4. demonstrate the statistics of weekly sharing of content/post by the survey participants either personal of educational.

e. Access to the Subject related content

In terms of subject related study/content, Facebook is providing great applications for both students and teachers and include access to documents, research material, and presentations. These applications include links, SlideShare, To-Do List, JSTOR Search, WorldCat, Zoho Online Office, Google Docs, Podclass and LibGuides. Therefore, important question in survey is to investigate that how much students access the material related to their subjects or field of study? In response of this question, it is found that only 39% students responded positively that they find their subject related material but remaining 61% participants declared that they do not find the content related to their subjects. The reasons may be lack of awareness, students' interest and teachers' support. In the figure no. 5. detailed responses of the survey participants are presented.

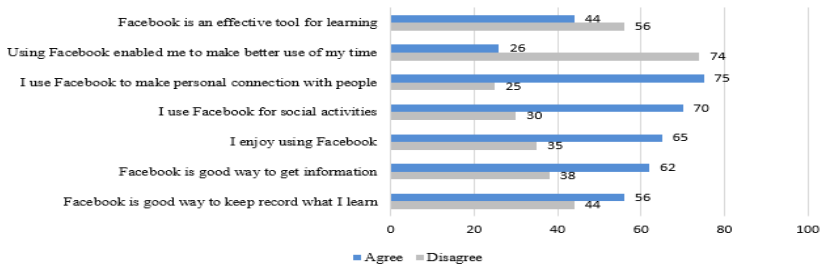
FIG. V. PRESENTS THE RESPONSES REGARDING CONTENT ACCESS



f. Usage priorites.

Currently Facebook is providing sophisticated resources for the teachers and students. The important applications for the students include study groups, campus buddy, Notecentric, skoolPool, class notes, Doresearch4Me and CiteMe etc. therefore most important part of this survey is to investigate the usage priorities of undergraduate students on the Facebook. These responses of survey participants have been shown in the figure no. 6, in the form of self-explanatory graph.

Fig. VI. GRAPH SHOWS THE USAGE PRIORITIES OF PARTICIPANTS IN PERCENTATE.



IV. DISCUSSION AND RECOMMENDATION

Social media has become integral part of the current generation. Although there is number of SNS like Facebook, Twitter, MySpace, Reunion, Babo and Friendster but Facebook has remained most commonly used social networking site. Although other SNSs have remarkably influenced every aspect of life including learning activities. Currently, every organization and institute etc. post their status via the SNS and share latest information and news through these social platforms. Facebook provides many facilities to its users like it is best platform for sharing information, provides secure and open discussion platform, individual and group chat, pages for marketing, social activities, updates and news, easy and secure sharing of data, voice and video calling facilities etc. Therefore, it is all in one social medium. These are the key attractive features of Facebook have attracted large community of the world. All these features can be very effective in order to augment learning either formal or informal. Moreover, availability of teachers will be very fruitful for the students in order to guidance and subject related discussion. Also teacher can monitor the activities, behavior and educational prorates of students.

In terms of research Facebook has become very ideal tool for research activities by providing the specific applications DoResearch4Me and CiteMe also it is good platform for survey, discussion, technical guidance and access latest information by joining interested research groups and specific pages the user can get lot of educational benefits. There are many research groups also which are benefiting to the scientific society. It is very ideal tool for the Awareness raising programs, keeping personal record and accessing current affairs. Therefore, it is very difficult to meet and discuss the problems with national and international subject experts, educationist, researchers, etc. but the SNSs make it not only possible but also more encouraging when user/student meet with these experts on this virtual and real environment, where large community share worth information and suggestions. In nutshell, attractive features of Facebook include easy access, interaction, information access to everyone, motivation, entertainment, multimedia support, information interpretation and provide a democratic environment. Along with these advantages there are some disadvantages also such as SNS are nested with entertainment, inaccurate information, informal convergence, information overload and lack of visual sharing.

In order to assure effective use of Facebook, teachers should motivate the students and guide them as students and get benefits from technological means of learning. Additionally, there are some suggestion also for positive usage and cooperative activities on Facebook.

- Attention is required regarding Facebook usage as an additional learning platform.
- Teacher-student interaction and removal of class hours' limitations.
- Training and group work about the educational use of Facebook or other SNSs.
- Increasing student participation by adding academic performance an assessment process.
- Supervision of teachers.
- Comments or involvement should be encouraged.
- Subject related sharing must be brief.
- Cooperative activities should be performed.

V. CONCLUSION.

In this paper we have presented literature review of important literacies of social media platforms specially Facebook. The existing literature suggests that Facebook is an E-learning environment, enhance and strengthen formal or informal learning. The interactivity and collaborative communication enhance cognitive and critical thinking skills. Specifically, attention, participation, collaboration, network awareness and critical consumption are the most important literacies of social media platforms. In the survey data, it is observed that majority of students use Facebook for social interaction with friends, family and also for news updates. However, very less number of students is involved in learning activities. The analysis of survey data reveals that access mediums have been also changed from traditional computers to smart phones. It is also observed that most of the participants remain less active in terms of sharing of content/posts or they just visit their Facebook accounts. It is investigated through this survey that most of the students are using Facebook as a traditional social interaction tool even at university level. Most of the features of Facebook have been unrealized by the majority of the students. However, some of the students assumed that, it could be used as an online environment for learning enhancement.

REFERENCES

[1] Rasiah, R. Ratneswary V. "Transformative higher education teaching and learning: Using social media in a team-based learning environment." Procedia-Social and Behavioral Sciences 123 (2014): 369-379.

[2] Wamba, Samuel Fosso, and Lemuria Carter. "Social media tools adoption and use by SMES: An empirical study." Journal of Organizational and End User Computing (JOEUC) 26.2 (2014): 1-17.

[3] Arshad, M., t al. "SOCIAL NETWORKING SITES: A PATH OF LEARNING IN HIGHER EDUCATION." Pakistan Journal of Science 66.4 (2014).

[4] Donlan, Leah. "Exploring the views of students on the use of Facebook in university teaching and learning." Journal of Further and Higher Education 38.4 (2014): 572-588.

[5] Evans, Chris. "Twitter for teaching: Can social media be used to enhance the process of learning?." British Journal of Educational Technology 45.5 (2014): 902-915.

[6] Dabbagh, Nada, and Anastasia Kitsantas. "Personal Learning Environments, social media, and self-regulated learning: A natural formula for connecting formal and informal learning." The Internet and higher education 15.1 (2012): 3-8.

[7] Cheung, Christy MK, Pui-Yee Chiu, and Matthew KO Lee. "Online social networks: Why do students use Facebook?." Computers in Human Behavior 27.4 (2011): 1337-1343.

[8] Yu, Angela Yan, et al. "Can learning be virtually boosted? An investigation of online social networking impacts." Computers & Education 55.4 (2010): 1494-1503.

[9] Cho, Hichang, et al. "Social networks, communication styles, and learning performance in a CSCL community." Computers & Education 49.2 (2007): 309-329.

[10] Cho, Hichang, et al. "Social networks, communication styles, and learning performance in a CSCL community." Computers & Education 49.2 (2007): 309-329.

[11] Liao, Yi-Wen, et al. "Exploring the antecedents of collaborative learning performance over social networking sites in a ubiquitous learning context." Computers in Human Behavior 43 (2015): 313-323.

[12] Ngai, Eric WT, Spencer SC Tao, and Karen KL Moon. "Social media research: Theories, constructs, and conceptual frameworks." International Journal of Information Management 35.1 (2015): 33-44.

[13] Ngai, Eric WT, et al. "Social media models, technologies, and applications: an academic review and case study." Industrial Management & Data Systems 115.5 (2015): 769-802.

[14] Lin, Jian-Wei, Hsieh-Hong Huang, and Yuh-Shy Chuang. "The impacts of network centrality and self-regulation on an e-learning environment with the support of social network awareness." British Journal of Educational Technology 46.1 (2015): 32-44.

[15] Powell, David V., et al. "Educational Technologies for Maturing Democratic Approaches to Educational Practices in Ukraine." Procedia-Social and Behavioral Sciences 176 (2015): 378-385.

[16] Nawaz, Asif, et al. "Impact of Social Media and Voice of Customer (VOC) in Decision Making Process an Empirical Study of Higher Education Institutions of Pakistan." Middle-East Journal of Scientific Research 23.6 (2015): 1069-1075.

[17] Wamba, Samuel Fosso, and Lemuria Carter. "Social media tools adoption and use by SMES: An empirical study." Journal of Organizational and End User Computing (JOEUC) 26.2 (2014): 1-17.

[18] Evans, Chris. "Twitter for teaching: Can social media be used to enhance the process of learning?." British Journal of Educational Technology 45.5 (2014): 902-915.

[19] Rasiah, R. Ratneswary V. "Transformative higher education teaching and learning: Using social media in a team-based learning environment." Procedia-Social and Behavioral Sciences 123 (2014): 369-379.

[20] MahboobRabbani, AnamJaveed, Muhammad Asif, Muhammad Ibrahim, "Impact of Social Networking Websites on Students learning", ISSN Vol.44, 2015

[21] Munienge Mbodila, Clever Ndebele and Kikunga Muhandji, "The Effect of Social Media on Student's Engagement and Collaboration in Higher Education", J Communication, 2014.

[22] Naveed Sultana, Social Media as Best Practices for Optimizing the Learning at Allama Iqbal Open University", Secondary Education Journal Volume 1, Issue 1 pp. 15-22.

[23] Chris Piotrowski, "Emerging research on social media use in education: a study of dissertations", Research in Higher Education Journal Volume 27, January 2015.

[24] Sameen Fatima1, Umar Manzoor2, Bassam Zafar2, and Mohammed A. Balubaid3, "Analyzing the Impact of Social Media on Users", IJCSI may 2015.

[25] Waleed Mugahed Al-rahmi, Mohd Shahizan Othman, Lizawati Mi Yusuf, "Using Social Media for Research: The Role of Interactivity, Collaborative Learning, and Engagement on the Performance of Students in Malaysian Post-Secondary Institutes" , ISSN Vol 6 No 5 S2September 2015

[26] Hira Hasnain, Anum Nasreen, Hamza Ijaz, " IMPACT OF SOCIAL MEDIA USAGE ON ACADEMIC PERFORMANCE OF UNIVERSITY STUDENTS", IRMIC 2015)

[27] Carmen Holotescu, Gabriela Grosseck "An Empirical Analysis Of The Educational Effects Of Social Media In Universities And Colleges", Volume 2 Issue 1 Spring 2013

[28] Paul A. Tess, "The role of social media in higher education classes (real a, nd virtual) – A literature review", 2013.

[29] Jeffrey Mingle, Dr. Musah Adams, "Social Media Network Participation and Academic.

[30] Mitch Wolf, Marianne, "Using Social Media to Enhance Learning through Collaboration in Higher Education: A Case Study" AAEA August 12 – 14, 2012

[31] Exploitation of social media among university students: A case study

[32] Gray, K., Annabell, L., & Kennedy, G. (2010). Medical students' use of Facebook to support learning: Insights from four case studies. Medical Teacher, 32(12), 971-976.

[33] Baran, Bahar. "Facebook as a formal instructional environment." British Journal of Educational Technology 41.6 (2010): E146-E149.

[34] Ali, A. (2016). Medical students' use of Facebook for educational purposes.Perspectives on medical education, 1-7.

[35] Pimmer, C., Linxen, S., & Gröhbiel, U. (2012). Facebook as a learning tool? A case study on the appropriation of social network sites from mobile phones in developing countries. British Journal of Educational Technology,43(5), 726-738.

[36] Gray, K., Annabell, L., & Kennedy, G. (2010). Medical students' use of Facebook to support learning: Insights from four case studies. Medical teacher, 32(12), 971-976.

[37] Cain, J., & Policastri, A. (2011). Using Facebook as an informal learning environment. American journal of pharmaceutical education, 75(10), 207.

[38] Pimmer, C., Linxen, S., & Gröhbiel, U. (2012). Facebook as a learning tool? A case study on the appropriation of social network sites from mobile phones in developing countries. British Journal of Educational Technology,43(5), 726-738.

[39] Alias, N., Siraj, S., Daud, M. K. A. M., & Hussin, Z. (2013). Effectiveness of Facebook based learning to enhance creativity among Islamic studies students by employing isman instructional design model. TOJET: The Turkish Online Journal of Educational Technology, 12(1).

A Proposed Framework for Testing Mobile Cloud Based Applications Using Mobile Testing as a Service (MTaaS)

Engr. Ali Ahmed
Computer & Software Engineering Department
Bahria University, Karachi Campus
Karachi, Pakistan
aliahmed.bukc@bahria.edu.pk

Huma Ali Ahmed
Department of Computer Science
Newports Institute of Communications & Economics
Karachi, Pakistan
huma.simran@gmail.com

***Abstract*—In this fast running growth of mobile cloud computing (MCC) and its development work grows day by day. Mobile cloud computing plays a dynamic role in technology for mobile devices. The MCC has been initiates as a reliable technology for the upcoming mobile services. MCC incorporates many tracks related to mobile environment and try to overcome the issues related to the security, performance, and architecture etc. This research paper gives a proper overview for mobile cloud applications development and testing issues. We are basically highlighting the issues of mobile cloud computing and try to propose some possible and reliable framework and testing methods with the help of mobile testing as a services (MTaaS) in this research paper. This framework can enhance the testing procedure for cloud based mobile applications.**

Keywords—*Mobile computing, Cloud computing, Mobile cloud computing, Mobile devices, Mobile cloud applications development, Mobile testing as a service*

I. INTRODUCTION

Mobile devices are now increasingly plays a major role in our daily life and becoming an essential need for us. It is now convenient and reliable communication medium for every person and it is not a bounded communication tool. Mobile users experience a lot of various services with the help of their mobile applications like Google apps which you can run on your devices very easily. Now there is a fast growing development of mobile computing and it plays a commanding development in the growing lives of information technology world. Some of the mobile devices face a lot of issues and challenges in their resources. Now a days, cloud computing gives some advantages for the users by their provided infrastructure. With the bang of mobile based applications and also the support of cloud computing, mobile users can integrate their devices over the cloud network. Cloud computing also facilitates the mobile users with the new types of services like gamming apps etc. This paper presents the overview on the upcoming trends and testing issues rise in MCC.

II. CLOUD COMPUTING

Through the Internet cloud computing platform gives the reliable resources and services. These services of cloud allow a single person or a business to use some kind of software and hardware that can be managed by another business. Cloud computing includes storage of files, many of the social networking sites, web mails, and all the online business applications. The advantage of cloud computing is that it allows the access to data information and the computer based reliable resources from anyplace to anywhere hence a network connection should be available [1]. Cloud computing model provides many kinds of resources, such as data storage, user applications, and networks. Cloud computing provides most of the developers and information technology departments with the ability on how to concentrate on which matters are the most important and to avoid things like procurement, maintenance, network speed and capacity. As cloud computing has now become popular, many of the different cloud services and the methods of deployment provides us with unusual levels that how to control, make it supple, and manage.

III. MOBILE COMPUTING

Mobile computing is basically a technology which allows the transmission of data from source to destination; also it transfers videos and voice etc. Through the help of any wireless device or computer without having to be connected with any physical link [2]. Some of the principles of mobile computing are stated below:

A. Mobile Portability

It provides the movements of devices from one place to another within the mobile computing environment.

B. Mobile Connectivity

In mobile connectivity it is able to stay continuously with the minimum amount of downtime without affected by movement of other devices.

C. Connecting Social Interactivity

It provides collaborative connectivity with the other users at least with the same environment.

D. Individuality

In the individuality phase it adapts the technology suite within their individual needs.

IV. Mobile Cloud Computing

MCC is a blend of mobile computing and cloud computing that also works for wireless networks. It works in a rich environment and for mobile users it provides computational resources to facilitate the customers, network operators and cloud computing providers [3]. In the track of mobile applications it is the new pillar for mobile apps where the data processing and storage are transfers from cell phones to resources. The basic structural design of MCC is illustrated in Fig. 1.

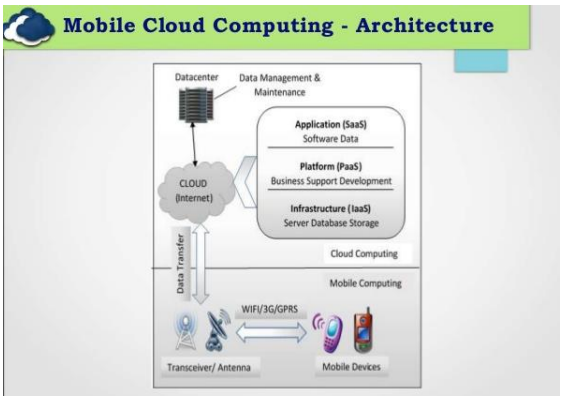


Fig. 1. Structural Design of Mobile Cloud Computing

A. Characteristics of Mobile Cloud Computing

Some of the mobile cloud computing characteristics are stated below:

1. Scalability
2. Reliability
3. Agility
4. Security
5. Reduced maintenance
6. Reduced cost

B. Challenges of Mobile Cloud Computing

Some of the mobile cloud computing challenges are stated below [4]:

1. Limitations of bandwidth and communication medium latency
2. Security issues
3. Signal disturbance and noise
4. Low computing backbone

Fig. 2 illustrates the comparison between testing procedure for cloud based MTaaS and traditional mobile testing [5].

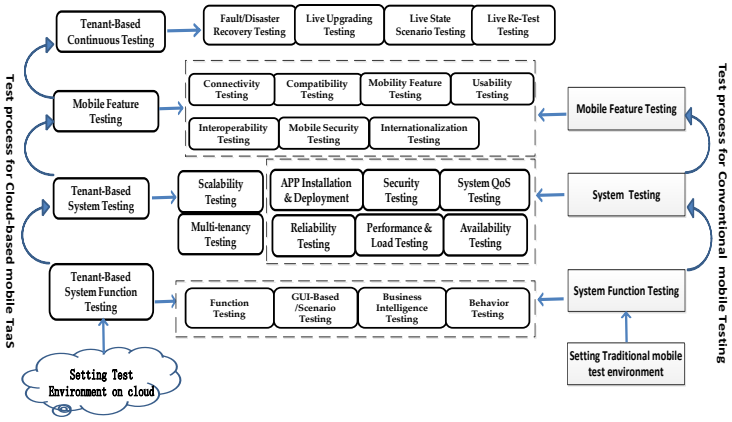


Fig. 2. Comparative Analysis of Process of Testing for Cloud Based MTaaS and Traditional Based Mobile Testing

Some of the tasks for cloud based mobile TaaS testing process are stated below:

A. Task 1

Set up the testing environment over the cloud, run and control test, monitor and tracking, interact with the server of TaaS as well as its essential mobile emulation cloud.

B. Task 2

In task 2, the system functional tenant based testing includes different types of tenant based features and functions.

C. Task 3

Scalability, multi-tenancy system checks, and quality of service (QoS) etc.

D. Task 4

It targets usability testing and mobility feature based testing etc.

E. Task 5

The validation of mobile systems including upgrading and recovery based on tenant based testing [6].

Cloud-based mobile TaaS has a variety of new testing environment, testing process and techniques etc. in comparison with conventional mobile testing as described in Table 1.

V. Mobile Testing As A Service

Mobile testing as a service provides on-demand testing service for mobile based apps and software as a service (SaaS) is to maintain the validation process and software quality. In quality based engineering process by leveraging a cloud based scalable and reliable architecture of mobile testing to assure the pre-defined service level agreements [7]. MTaaS offers novel big business framework for different mobile based validation services for software by using the pay-as-you-testing model to accomplish the goal of cost reduction and cost sharing in the resources of mobile based computing cloud based computing, networking and the structure of storage.

A. Requirements and Needs for MTaaS

Some of the requirements and needs for MTaaS are stated below:

1. Mobile based on demand testing
2. Easy connectivity and mobility
3. Configuration and control of mobile testing infrastructures
4. Customization support
5. Monitoring and tracking
6. Coverage analysis

	Conventional Mobile Based Testing	Cloud-Based Mobile Testing as a Service
Primary Objectives	Authenticate the excellence of mobile applications on mobile operation environments of specified mobile devices or from different web browsers.	It provides on-demand testing services on cloud; leverage a scalable cloud based mobile testing environment to assure pre-defined given service-level agreements (SLAs) and QoS.
Testing Focuses	Diverse software errors in its structures, functions, behaviors, user interfaces, and connections to the external systems. System non-functional requirements such as reliability, scalability and security.	Customization, multi-tenancy, and configurability. SaaS scalability (including both horizontal and vertical scalability). Connectivity to its external contexts, interoperability and portability.
Execution of Testing	Offline testing must be done in a test lab before the product deployment.	In private testing environment the major chunk is offline testing; we need on-demand test execution in a virtual cloud based test environment.
Testing Environment	In the testing environment a pre configured test lab is available with purchased hardware/software and tools.	A virtual scalable mobile test environment based on cloud environment with diverse computing resources and tools; supporting web browsers on different mobile platforms and devices.
Testing Process	Enterprise-oriented test processes for each project.	Crowd sourcing-based process; well-defined TaaS processes by TaaS vendors; emphasis on tenant-based testing, service component testing, etc.
Testing Techniques	Apply black box and white box well known testing techniques at unit level or at the system level.	It needs original continuous techniques of testing. Latest solutions for testing is to compact with multi- tenancy.
Testing Tools	Use limited testing solutions and tools with the purchased licenses.	On-demand usage of diverse test tools with shared licenses in a cloud environment based on pay-as-you-use.
Project Cost	Required hardware/software (license) costs in a test lab, plus the cost of engineering.	It is based on a pay-as-you-test service costs pre-defined SLA.

Table 1. A Comparison of Conventional Mobile Testing and Cloud-Based Mobile TaaS

B. Various Infrastructures of MTaaS

In MTaaS we have three types of infrastructures; first is emulation based testing on cloud, second is simulation based testing on cloud and last is device based mobile testing on cloud.

1. Emulation Based Testing on Cloud

In emulation based testing the mobile based SaaS instances which are on a cloud validate themselves by using large scale mobile emulation over the cloud. Testing approach in emulation based testing is usually contains inexpensive approaches because we don't need real mobile devices. At the end this approach has flaws and limitations in testing.

2. Simulation Based Testing on Cloud

Simulation based testing provides mobile application servers which are available on cloud and validate themselves by using mobile simulators over the cloud. But this approach never requires real mobile devices. On the hand other it has some difficulties in validating few features of system.

3. Device Based Mobile Testing on Cloud

In device based testing real devices needs to be purchased that will use to validate the software apps like mobile based SaaS and web applications [8]. Well configured mobile devices set up to meet the mobile based testing services according to the requirements of on-demand test service. Various mobile TaaS infrastructures with models are discussed in Table 2.

Approaches / Perspectives	Emulation Based Testing on Clouds	Mobile Testing in Crowd Sourcing	Device Cloud Based Mobile Testing
Mobile TaaS Service Model	Service model for emulation based mobile TaaS	Service model for crowd based sourcing	Model for remote mobile based TaaS service
Billing and Business Models	Pay-as-you-use for device emulators and other testing services	Crowd sourcing cost models	Pay-as-you-use for remote devices and other testing services
Mobile Testing Environment	Emulation-based mobile TaaS infrastructure	Environment of ad-hoc mobile testing, and TaaS architecture	Shared mobile TaaS infrastructure
Mobile Devices Costs	Only use emulators, no device costs	No costs on mobile devices since it uses mobile devices from crowded testers	Device rental costs
Mobile End-To-End Transaction Testing	Emulation-based end-to-end transaction testing	End-to-end transaction testing in crowd sourcing	Large-scale device oriented end-to-end transaction testing
Mobile Usability Testing	No coverage on real mobile user experience	Easy to support usability testing	Scalable test coverage on real mobile devices
Mobility and Location-Based Testing	No reallocation service testing, using emulators only	Easy to perform location service testing by crowded testers	Limited location service testing
QoS Testing for Performance, Reliability and Availability	Emulation-based QoS testing at the limited scale	Ad-hoc QoS testing using crowd sourcing, at the risk of low testing quality and an uncertain validation schedule	Large-scale QoS testing
Mobile Security Testing	Emulation-based security testing only	Ad-hoc security testing only in crowd sourcing	Diverse device based security testing
Scalable GUI Testing	Emulation-based, limited scale	Scalable GUI testing based on number of users	Large scale GUI testing on real devices
Mobile App Function Testing	Emulation-based function testing, hard to test functions related to real devices	Ad-hoc mobile function testing	Diverse device based function testing

Table 2. An Evaluation Assessment between Cloud-Based MTaaS Perspectives

VI. Proposed Framework For Testing Mobile Cloud Based Applications Using Mobile Testing As A Service

This proposed framework is a mixture of cloud computing and mobile cloud computing. Basically we have designed this model for testing cloud based mobile applications using MTaaS. MTaaS provides on-demand testing in its interface for mobile based applications usually. In this model, we have proposed the testing environment for the testers so they can easily test the applications with the help of this framework Even they can maintain the test cases and also with the help of this framework a tester can make the reports and rectify the bugs which can harm the applications. Fig. 3 illustrates the said concept.

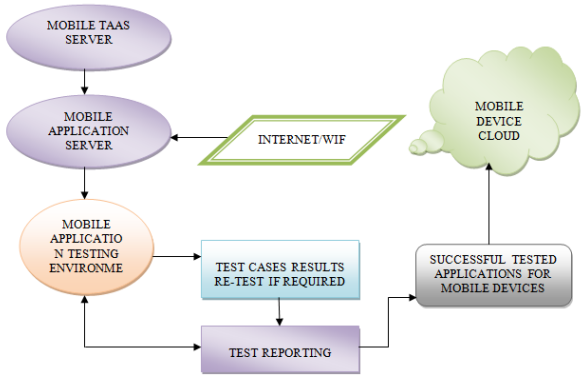


Fig. 3. Framework for Testing and Updating Cloud Based Mobile Applications

VII. Conclusion And Recommendations

This research paper proposed that the mobile cloud computing and cloud computing are dependent on each other; they can work together and provide scalable resources. We use MTaaS which offers on-demand testing services for mobile applications. With the help of this framework, we can facilitate the cloud based mobile applications by making test plans and especially we can generate the test reports. For future work, we can add more reliable resources in the architecture of this model and we can add more features which can help this framework to work well and become easy for cloud based mobile applications testing.

References

[1] S. Kumar and R. H. Goudar, "Cloud Computing – Research Issues, Challenges, Architecture, Platforms and Applications: A Survey," International Journal of Future Computer and Communication, vol. 1, no. 4, pp. 356-360, 2012.

[2] X. Peng and Y. Guofeng, "DCSP-MC: Dependable Cloud-Based Storage Platform for Mobile Computing," International Journal of Networking and Virtual Organisations, vol. 12, no. 2, pp. 133-148, 2013.

[3] E. Ibukun and O. Daramola, "A Systematic Literature Review of Mobile Cloud Computing," International Journal of Multimedia and Ubiquitous Engineering, vol. 10, no. 12, pp. 135-152, 2015.

[4] M. R. Sudha and I. V. Lema, "A Study on Emerging Trends and Challenges in Mobile Cloud Computing," International Journal on Recent and Innovation Trends in Computing and Communication, vol. 4, no. 1, pp. 234-237, 2016.

[5] R. Temkar, S. Gadekar and D. Shah, "Cloud Based Mobile Application Testing," International Journal of Science, Engineering and Technology Research, vol. 4, no. 6, pp. 2097- 2102, 2015.

[6] A. C. Donald, S. A. Oli and L. Arockiam, "Mobile Cloud Security Issues and Challenges: A Perspective," International Journal of Engineering and Innovative Technology, vol. 3, no. 1, pp. 401-406, 2013.

[7] J. Gao, W. T. Tsai and R. Paul, "Mobile Testing-as-a-Service (MTaaS) -- Infrastructures, Issues, Solutions and Needs," High-Assurance Systems Engineering, IEEE 15th International Symposium, pp. 158-167, 2014.

[8] K.Vivekanandan, R. Bhojan and S. Ganesan, "Cloud Enabled Test Evaluation on Mobile Web Applications," International Journal of Advanced Research in Computer and Communication Engineering, vol. 3, no. 6, pp. 6933-6937, 2014.

Malleability issues in Block-Chain applications

Kashif Mehboob Khan
Department of Computer Science & Information Technology, NED University of Engineering & Technology-Karachi, Pakistan
kashifmehboob@neduet.edu.pk

Muhammad Mubashir Khan
Department of Computer Science & Information Technology, NED University of Engineering & Technology-Karachi, Pakistan
mmkhan@neduet.edu.pk

***Abstract*— The block-chain mechanism is being implemented in diversified areas of real world applications especially in financial transactions. It has been observed that these transactions may induce malleability in a number of ways to the block chain and do have the tendency to produce the problem of double utilization of same token. Such problems may cause some real time threats to the real world systems if certain conditions are met. For instance, it may cause a genuine owner to be illegally deprived of his assets when a malicious user intentionally discontinues further propagation of his block containing the transaction that moves token from his address to the seller's address. In this paper we propose an attack model to show the double utilization of same token values in block-chain design. The paper concludes by highlighting possible countermeasures of double spending problem.**

Keywords*— *block-chain; malleability; risks; malacious transactions.

I. INTRODUCTION

The block-chain system has been designed for transaction databases (usually financial transactions), which are publicly shared by all the nodes in the network. Every transaction contains a transaction id which is actually hash of all the fields in a transaction. Now it has been known since roughly 2011 that signed transactions are slightly "malleable" in the sense that it is possible to modify a signed transaction in certain minor ways, without invalidating the signature [1][2][11]. Cryptography ensures that the critical details about a transaction can not be changed (like sender, receiver, amount etc.) but certain non-functional fields that do not contribute to the critical parts of a transaction may be changed which causes the hash (transaction id) to be changed for the same transaction.

When transactions in a block-chain are signed, all the data in a transaction is not covered in the creation of transaction hash which makes it possible for an attacker on the block-chain network to change the transaction in such a way that the hash is nullified. This changes the hash of the transaction only, while the output and the message of the transaction remains same. Therefore, in order to avoid transaction malleability one should not accept the transactions that are not mined or confirmed, because all the following transactions in a block-chain depends upon the hashes of the previous transactions, and those hashes can be changed until they are confirmed in a block. Double spending is the possibility to spend a transaction twice or more claiming the same input as a consequence of transaction malleability. One of the transactions will be included in the public ledger while other will be discarded by the network as it will be considered invalid.

One way to check the malleability impact in Bitcoins, is to artificially inject multiple malleable transactions immediately after an original transaction by just changing the nonfunctional fields of a transaction so that new hashes (transaction ID's) may be formed against a practically same transaction. Now if any one of the malleable transactions gets mined first before the original transaction, the miners (computers in the network which validate the transactions) will add this transaction to the block as a valid one because the critical fields in the transaction were unchanged. Now if the sender of the transaction looks for the confirmation of transaction by its transaction id in the transaction database (publicly shared blocks), he is never going to find it as the original transaction (which in our case could not be mined first) would be rejected by the miner as a double spent [2][3]. Malleability may affect badly to other applications based on the data structure of block-chain. For instance, the same person may cast multiple votes.

This paper is organized in three sections. Section 1 introduces the problem and discuss some of the existing approaches which are used to tackle malleability. Section 2 provides references to the work related to the transaction malleability in block chain. In Section 3, an attack model is presented, which is supported by a practical example scenario to show a malleable transaction. The last part of this section shows the experimental working where a block-chain based network model has been formed using multi-chain as a platform. The future work will be based upon applying different methods of injecting malleable transaction in the above mentioned block-chain model and observe its impacts technically and socially on various commonly used applications which can make use of block-chain data structure.

II. RELATED WORK

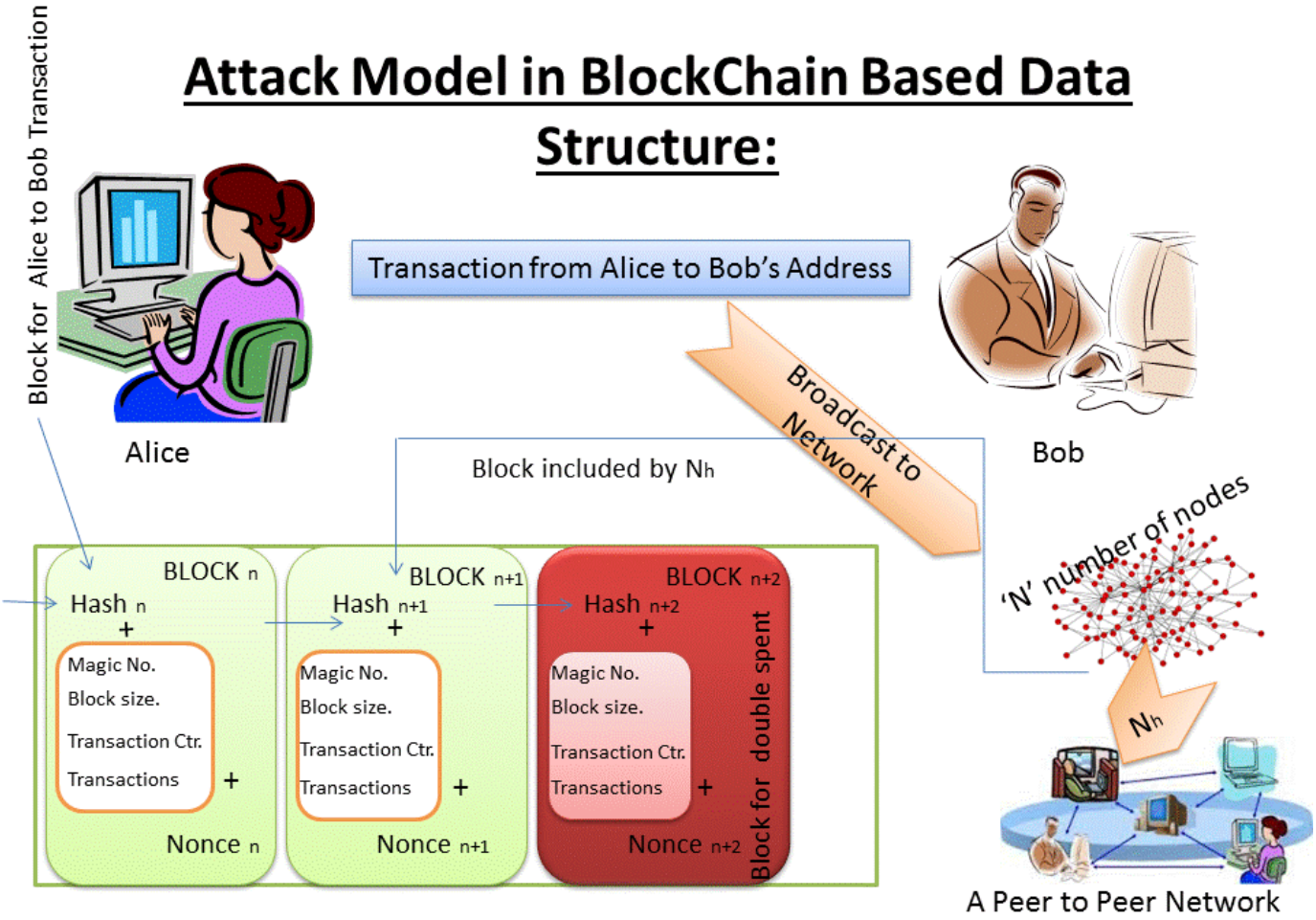
Several approaches have been used out to minimize the risks of double spent transactions which make use of malleability. Among such approaches, one is to wait for certain confirmations which are usually from six miners, to validate a transaction [4]. Even then, it cannot be guaranteed

that the true branch of block chain will be proceeded by true miners as due to network latency, there is a chance that the transaction which has occurred later may be listened earlier than the actual transaction. Therefore, the ordering of events which is based upon time stamps, does not show the true state of the system. There are also some other ways where researchers have made efforts to resolve the problem of double spend, but these are mostly by forcing certain constraints to keep lock of transaction for a particular duration of time until it gets completed. This has been explained by Kadam et al. in their paper [3, 4]. The problem that may arise here is that there is no global time, therefore the area is still open for research community to explore [4].

III. THE ATTACK MODEL

and contains a transaction from sender "S" to receiver "R". A transaction's data structure contains sender's signature, an instruction to send token to receiver's public key, and a hash. This hash is a pointer to a previous transaction output that the sender "S" received and is now spending. That pointer must reference a transaction which was included in some previous block in some consensus chain.

When the receiver finds his transaction included in the consensus block, the sender is acknowledged. Now consider a case in which the next selected node "N" happens to be controlled by the sender "S", then there is a very fair chance that this new node "N" ignores the block that contains the transaction from "S" to "R" in which "S" has moved token from its address to receiver's address and adds new block prior to that block. Moreover, the new proposed block may



In order to conduct test for malleability, an attack model may be developed to operate on the system. Let's assume that a sender 'S' moves his token from his address to receiver "R" on his address against the product or service he wants to utilize. Suppose the transaction successfully occurs from "S" to "R" and it is confirmed by an honest node and gets this transaction into the block chain. So the situation is that there is a block in the block chain which is added by an honest node also contain a transaction that moves values from sender account "S" to another account which is also controlled by "S", thereby generating opportunity to reuse the same token twice [4, 6].

A. *Modelling and Assumptions*

The above mentioned scenario for double spending may be modelled this way. When a double spend attack is made, the

network is in a position that contains a branch which moves the token to the vendor and has *n* blocks extending the one where fork started. In order to model the above mentioned scenario, we can make following assumptions;

i. The system is cryptographically secured and our scope includes how to defend against double utilization of the same token (double spend).

ii. Let the combined hash rate of honest network and the attacker is constant and denoted by *H* and pH belongs to the hash rate of honest network and qH belongs to that of attacker, where p+q=1

iii. Also assume that mining difficulty is constant. *To* is the average time to find a block with a hash rate of H

Let *z* represent the added number of blocks which are created by honest miners from the block where the fork started and has an advantage over miner;

Mathematically,

$$z = n - m$$

where n is the total number of blocks which are created after the transaction that is transferred to vendor, m is the total number $z = n - m$ of blocks which are created by miner (attacker).

The value of z increments or decrements by 1 if the block is added by honest network and miner (attacker) respectively. Here it becomes a continuous–time Markov chain where P/T_o and Q/T_o for increasing and decreasing the chain respectively. Note that if at any stage, the value of *z* becomes negative, it is obvious that the attacker's chain is now bigger and therefore the attack is successful. In order to find whether z will ever be -1, we can take the help of discrete time Markov chain process where the step of the process is defined as the finding of block either by honest network or miner (attacker). P is the probability that the block is found by honest network and *q* be the probability that the block is found by the attacker [2][3][17].

Let a_z shows the probability that the attacker will be succeeded when he is z blocks behind. Now if z is negative then a_z approaches to 100% as he will have a longer branch than the honest miner. Assume that the next block is found by the honest network, which happens with probability p, the attacker will now be $z + 1$ blocks behind and his probability of success will be a_{z+1}. If the next block found will be by the attacker, which happens with probability q, his probability of success will be a_{z-1} [2][14][6].

$$a_z = pa_{z+1} + qa_{z-1}$$

B. Tools & Technologies

Following tools and technologies were considered and explored for creating block-chains and mining of blocks:

- **DESMO-J** is a framework for Discrete-Event Modelling and Simulation which is built on Java. It supports both the process oriented and event oriented modelling style, also known as process interaction approach and event scheduling approach respectively. DESMO-J library can be configured to test and verify the simulation results of discrete events which has been used by many researchers in their research. Obviously, It does not run over real network as it is just a java based library for processing discrete events.

- **Confidence Chains**, a project developed specifically for bitcoin, offers a very flexible way to define the trust relationship that is suitable for a wide range of applications. It also offers a very high degree of irreversibility that does not necessarily depend on the direct authority of one party It has similar anonymity and security characteristics to bitcoins. However, to the best of our knowledge, it was not developed to run over network and incorporate block-chain based applications in general.

- **Multi-Chain** is an off the shelf platform for the creation and deployment of private block-chains either within or between organizations. It aims to overcome a key obstacle to the deployment of block-chain technology in the institutional financial sector by providing the privacy and control required in an easy to use package. Like the bitcoin core software from which it is derived Multi-Chain supports Windows, Linux and Mac servers and provides a simple API's interface and command therefore it was selected as a final platform for building the real network based architecture to perform transactions through block-chain among nodes

C. Implementation of Block-Chain based Network Architecture

The experiment was desired to be conducted by utilizing block-chain in a scenario other than bitcoin to keep focus on block-chain data structure which is the basic engine behind all the transactions and all the issues associated with it. In this example block-chain based scenario has been discussed. Some terminologies necessary to understand the scenario are as follows [21]:

- E-Voting refers here to the block-chain based electronic voting.

- E-Voting Participants include voter, registration authority and candidates.

Following are the steps to build Multi-Chain

i. Configuring Block-Chain for E-Voting

```
>>kashif@kashif-pc:~/.multichain$ mkdir node1
>> kashif@kashif-pc:~/.multichain$ multichain-util
create voteChain
Multichain utilities build 1.0 alpha 16 protocol 1003

Blockchain parameter set was successfully generated.

You can edit it in
/home/kashif/.multichain/voteChain/params.dat before
```

```
running multichaind for the first time.

To generate blockchain please run "multichaind
voteChain".
```

Fig_01

ii. Starting the Block-Chain through root node.

```
>>kashif@kashif-pc:~/.multichain/voteChain$
multichaind voteChain -daemon -
datadir=/home/kashif/.multichain/voteChain


Multichain utilities build 1.0 alpha 16
protocol 1003

Multichain server starting

kashif@kashif-pc:~/.multichain/voteChain$
Looking for genesis block…..
Genisis block found
New users can connect to this node using
multichaind voteChain@192.168.1.107:7721

Node started
```

Fig_02

iii. Creating second node.

```
>>kashif@kashif-pc:~/.multichain$ mkdir
voteChainnode2
>>kashif@kashif-pc:~/.multichain$ multichaind
voteChain @192.168. 1.107:7721 -
datadir=/home/kashif/.multichain/voteChainnode2
-rpcport=6001 -port6002

Multichain utilities build 1.0 alpha 16
protocol 1003

Retrieving blockchain parameters from the seed
node 192.168.1.107:7721 ...

New users can connect to this node using
multichaind voteChain@192.168.1.107:6002

Node started
```

Fig_03

iv. Retrieving chain info.

```
>>voteChain : getinfo {"method" : "getinfo","params"
:[],"id":1,"chain_name"::voteChain"}

{

      "version" : "1.0 alpha 16",
      "protocolversion" : 1003,
      "chainname" : "voteChain",
      "description" : "Blockchain for voting",
      "protocol" ; "multichain",
      "port" : 7721",
       …… ,
       ……..,
       ……..,

 }
```

Fig_04

v. Generation Of Voter Addresses From Node's Wallet

After creating nodes, multiple addresses were created from each wallet.

```
>>voteChain: getnewaddress
{"method" : "getnewaddress","params"
:[],"id":1,"chain_name"::voteChain"}


1QoVDR6Qkexwsop6f6d204d756c7469436861696eDVk2
```

Fig_05

vi. Issuance of Assets to Addresses

After successful generation of addresses, assets (representing votes) were created and assigned to each address.

```
>>voteChain: issue
1QoVDR6Qkexwsop6f6d204d756c7469436861696eDVk2 vote1 1
1 {"method" : "issue","params" :["
1QoVDR6Qkexwsop6f6d204d756c7469436861696eDVk2"],"id":1
,"chain_name"::voteChain"}


dc05e2c6Qkexwsop6f6d204d756c7469436861abdc6089
{"method" : "listassets","params"
:[],"id":1,"chain_name"::voteChain"}

[

   {

      "name" : "vote1",
      "issuetxid" : "
dc05e2c6Qkexwsop6f6d204d756c7469436861abdc6089",
      "assetref" : "70-265-1500",
      "multiple" : 1,
      "units" : 1,
       "details" : {
       },
       "issueqty" : 1.00000
       :issueraw" : 1

    }

]
```

Fig_06

vii. Performing Transactions between Addresses

A transaction was performed between different nodes.

```
>>voteChain: sendassettoaddress
1HkCKXXXXXXXXXXXXXXp7F vote1 1 {"method" : "
sendassettoaddress","params" :["
1HkCKXXXXXXXXXXXXXXp7F","vote1"]
,"id":1,"chain name"::voteChain"}
b971da6b60fXXXXXXXXXXXXXXXX43481cf
{
{"method" : "getaddressbalances","params" :["
1HkCKXXXXXXXXXXXXXXp7F","vote1"]
,"chain name"::voteChain"}

[
    {
     "name" :  vote1",
      "assetref" : "70-265-1500"
       "qty"  :  1.0000
    }
```

Fig_07

TABLE 1

Platform	Development Technology		
	Client side	***Server side***	***Database***
Ubuntu	Multi-chain Client	Multi-chain Server	Block-chain

IV. CONCLUSION AND FUTURE WORK

Since the block-chaining mechanism has been implemented successfully using Multi-Chain platform and commands using Ubuntu Shell and it has been observed that block-chaining mechanism is not restricted to bitcoin only rather it may be applied on many other diversified application like 'E-voting', therefore the risks of malleability which is mostly associated with bitcoin transactions is in-fact a block-chain oriented problem and needs to be addressed independently of the application in which block-chain based data structure is being utilized. Our future work will target the implementation of above mentioned attack model with reference to E-Voting to observe and test the system over all behavior and its potential impact on the factors associated with the conventional voting system.

ACKNOWLEDGMENT

The research work is supported by N.E.D. University of Engineering & Technology.

REFERENCES

[1] Daniel Kraft, "Difficulty Control for Blockchain-Based Consensus System", Peer-to-Peer Networking and Applications by Springer, March 2015.

[2] M. Rosenfeld. "Analysis of hashrate-based double-spending." [Online]. Available: http://arxiv.org/abs/1402.2009

[3] Mandar Kadam, Praharsh Jha, Shravan Jaiswal, "Double Spending Prevention in Bitcoins Network", International Journal of Computer Engineering and Applications, August 2015.

[4] S. Nakamoto. (2009) Bitcoin: "A peer-to-peer electronic cash system". [Online]. Available: http://bitcoins.info/bitcoin-a-peer-to-peer-electroniccash-system-satoshi-nakamoto

[5] G.O. Karame, E. Androulaki, and S. Capkun. Two bitcoins at the price of one doublespending attacks on fast payments in bitcoin. In Proc. of Conference on Computer and Communication Security, 2012

[6] Arvind Narayanan, Joseph Bonneau, Edward Felten, Andrew Miller, Steven Gold feder, "Bitcoin and Cryptocurrency Technologies", Chapter 2 and 3,Draft Oct0ber 2015

[7] J. Gobel, H.P. Keeler, A.E. Krzesinski, P.G. Taylor, "Bitcoin Blockchain Dynamics: the Selfish-Mine Strategy in the Presence of Propagation Delay", May 2015

[8] Guy Zyskind, Oz Nathan, Alex 'Sandy' Pentland, "Decentralizing Privacy: Using Blockchain to Protect Personal Data" IEEE CS Security and Privacy Workshops, 2015

[9] C. Decker and R. Wattenhofer, "Information propagation in the bitcoin network," in 13th IEEE Conference on Peer-to-Peer Computing, 2013, pp. 1–10

[10] C. Decker and R. Wattenhofer, "Information propagation in the bitcoin network," in 13th IEEE Conference on Peer-to-Peer Computing, 2013, pp. 1–10

[11] L. Bahack, "Theoretical bitcoin attacks with less than half of the computational power." 2013 [Online]. Available: eprint.iacr.org/2013/868

[12] F. Baccelli, I. Norros, and F. Mathieu, "Performance of p2p networks with spatial interactions of peers." [Online]. Available: http://hal.inria.fr/inria-00615523v2

[13] George Foroglou, Anna-Lali Tsilidou, "Further Applications of Blockchain" 12th Student Conference on Managerial Science and Technology, At Athens ,Conference Paper, May 2015

[14] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. "l-diversity: Privacy beyond kanonymity". ACM Transactions on Knowledge Discovery from Data (TKDD), 1(1):3, 2007.

[15] Petar Maymounkov and David Mazieres. Kademlia, "A peer-to-peer information system based on the xor metric", In Peer-to-Peer Systems, pages 53–65. Springer, 2002

[16] Arvind Narayanan and Vitaly Shmatikov, "How to break anonymity of the netflix prize dataset".arXiv preprint cs/0610105, 2006.

[17] Latanya Sweeney. k-anonymity, "A model for protecting privacy", International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05):557–570, 2002.

[18] Antonopoulos, A.M, Mastering Bitcoin: Unlocking Digital Cryptocurrencies. O'Reilly Media, 2014

[19] Ittay Eyal and Emin Gun Sirer, "Majority is not Enough: Bitcoin Mining is Vulnerable." ,18th International Conference on Financial Cryptography and Data Security. Barbados, 2014

[20] Swartz, A.: Squaring the Triangle: Secure, Decentralized, Human-Readable Names. January 6th, 2011

[21] http://www.multichain.com/developers/json-rpc-api/

Analyzing Diabetes datasets Using Data Mining Tools

Anita Shaikh
Department of Computer Science and Information Technology,
NED University,
Karachi, Pakistan.
aneeta_2013@live.com

Sohail Abdul Sattar
Department of Computer Science and Information Technology,
NED University,
Karachi, Pakistan.
sattar@neduet.edu.pk

Warda Tariq
Department of Computer Science and Information Technology,
NED University,
Karachi, Pakistan.
warda_tariq@rocketmail.com

Anaum Hamid
Department of Computer Science and Information Technology,
NED University,
Karachi, Pakistan.
anaumhamid@gmail.com

***Abstract*— Analyze, examine, explore and to make use of data this we termed as data mining .Data mining is useful in various fields for eg in medicine and we may take help for predicting the non-communicable diseases like diabetics. Diabetes mellitus placed 4th among NCDs, caused 1.5 million global deaths each year worldwide [1]. We are using different classifying algorithms such as Naïve bayes , MLP, J.48, ZeroR, Random Forest, Regression to depict the result and compare them and our aim is to find solution to diagnose the disease by getting meaningful result out of the data**

Keywords—Data mining; Classification; Algorithm;Diabtes MelitusType II.

I. INTRODUCTION

Non-communicable diseases (NCDs) which include stroke, heart disease, cancer, chronic lung cancer and diabetes they together are responsible for almost 70% of the deaths worldwide [1] in which Diabetes mellitus Type ii is most common in all.

The number of patients suffered has quadrupled since 1980.it is estimated that 422 million people have diabetes all over the world and this figure may get doubles in the next 20 years. [1]

The top 10 countries which are affected are India, China, USA, Indonesia, Japan, Pakistan, Russia, Brazil, Italy and Bangladesh.[3]

About seven million Pakistanis had diagnosed Type ii diabetes mellitus it is estimated that in 2035, the figure will reach up to 12 million. [4]

In this situation, we need to look into the facts and figure and the risk factors involved in it.

This paper meant to be written to give us an idea so we can early on diagnose the disease by comparing different data mining algorithm.

II. BACKGROUND

Data mining is a process in which we are allowed to ascertain the patterns in the provided datasets by simply applying combination of methods like artificial intelligence, machine learning, statistics and database system.

Our objective is to obtain information from the dataset and alter it to a more meaningful structure that is understandable to a layman.

The data mining tool that we are opting for is WEKA. It gives us the ideas as it contains tools for data pre-processing, classification, regression, clustering, association rules and visualization. It is also suited for developing new machine learning schemes.[2]

In this particular example, we are using different classifiers which include naïve bayes, decision tree and regression techniques and neural networks to get the best results out of it.

III. METHODS

A. . Classification

Naiive Bayes
ZeroR
MLPUnits

B. Regression

Logistic Regression

C. Decision Tree

J48
Random Forest

IV. DATASETS

The datasets had been taken from Pima Indians Diabetes Database of National Institute of Diabetes and Digestive and Kidney Diseases these datasets includes records of 768 patients, out of which 500 tested negative while 268 of them were tested positive.[9]

In this, we are analyzing the data with these nine Attributes that helps us to understand the possible prediction of this disease and which of the algorithm is more suitable for it.

The first eight attribute are the inputs set as input and the ninth attribute is the result which is used as a target which either “Positive” or “Negative”.

TABLE I. DATASETS OF DIABETIC PATIENTS

S.NO	Name	Description	Unit	Value range
01	Preg	No of Times Pregnant	Numeric value	0-9
02	Plas	Plasma GulocoseConcentrartion	Numeric value	0-199
03	Press	Diastolic Blood Pressure	mmHg	0-122
04	Skin	Triceps skin folds thickness	mm	0-99
05	Insulin	2-Hours Serum Insulin	mu/Uml	0-846
06	Mass	Body Mass Index	Weight in kg Height in m^2	0-67.1
07	Pedi	Diabets Pedigree Function	Numeric value	0.08-2.42
08	Age	Age	Numeric value	21-81
09	Classs	Diabetes Melitis Type II	Numeric value	Postive =1 ,Negative = 0

TABLE II. TABLULAR VIEW OF DATASETS

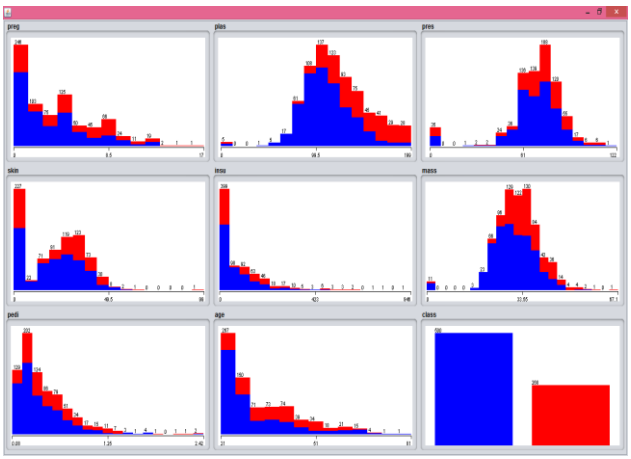
	preg	plas	pres	skin	insu	mass	pedi	age	class
1	6	148	72	35	0	33.6	0.627	50	tested_positive
2	1	85	66	29	0	26.6	0.351	31	tested_negative
3	8	183	64	0	0	23.3	0.672	32	tested_positive
4	1	89	66	23	94	28.1	0.167	21	tested_negative
5	0	137	40	35	168	43.1	2.288	33	tested_positive
6	5	116	74	0	0	25.6	0.201	30	tested_negative
7	3	78	50	32	88	31	0.248	26	tested_positive
8	10	115	0	0	0	35.3	0.134	29	tested_negative
9	2	197	70	45	543	30.5	0.158	53	tested_positive
10	8	125	96	0	0	0	0.232	54	tested_positive
11	4	110	92	0	0	37.6	0.191	30	tested_negative
12	10	168	74	0	0	38	0.537	34	tested_positive
13	10	139	80	0	0	27.1	1.441	57	tested_negative
14	1	189	60	23	846	30.1	0.398	59	tested_positive
15	5	166	72	19	175	25.8	0.587	51	tested_positive
16	7	100	0	0	0	30	0.484	32	tested_positive
17	0	118	84	47	230	45.8	0.551	31	tested_positive
18	7	107	74	0	0	29.6	0.254	31	tested_positive
19	1	103	30	38	83	43.3	0.183	33	tested_negative
20	1	115	70	30	96	34.6	0.529	32	tested_positive
21	3	126	88	41	235	39.3	0.704	27	tested_negative
22	8	99	84	0	0	35.4	0.388	50	tested_negative
23	7	196	90	0	0	39.8	0.451	41	tested_positive
24	9	119	80	35	0	29	0.263	29	tested_positive
25	11	143	94	33	146	36.6	0.254	51	tested_positive
26	10	125	70	26	115	31.1	0.205	41	tested_positive
27	7	147	76	0	0	39.4	0.257	43	tested_positive
28	1	97	66	15	140	23.2	0.487	22	tested_negative
29	13	145	82	19	110	22.2	0.245	57	tested_negative
30	5	117	92	0	0	34.1	0.337	38	tested_negative
31	5	109	75	26	0	36	0.546	60	tested_negative
32	3	158	76	36	245	31.6	0.851	28	tested_positive

V. GRAPHICAL REPRESENTATION OF ATTRIBUTES

Test Positive= Red

Test Negative =Blue

Fig. 1. Decision Tree (Random Forest)



Classification Algorithm and Their Evaluation

Output Prediction

These results were based on 90% percentage split In this table the column "Actual " and "predicted" represents the original result versus the predicted result however the column "error" represents the prediction error.

TABLE III. COMPARISON OF PREDICTION OF FIRST TWO INSTANCES BY USING DIFFERENT ALGORITHMS

N= tested_negative
P= tested_positive

Decision Attributes	Logistic Regression	Naïve Bayes	ZeroR	J.48	MLP	Random Foest
Instance 1						
Actual	N	N	N	N	N	N
Predicted	N	N	N	N	N	N
Prediction (True/False)	True	True	True	True	True	True
Instance 2						
Actual	P	P	P	P	P	P
Predicted	P	N	N	P	N	N
Prediction (True/False)	True	False	False	True	False	False

A. Naïve Bayes

This algorithm is named after Thomas Bayes who proved the bayes theorem. Naive Bayes is suitable in our situation in which we need to identify the possibilities of how many people are more prone towards diabetes.
This algorithm works on probability distribution function.

TABLE IV. PREDICTION TABLE

inst#	actual	predicted	error
1	1:tested_negative	1:tested_negative	0.99
2	2:tested_positive	1:tested_negative	+0.67
3	1:tested_negative	1:tested_negative	0.501
4	1:tested_negative	1:tested_negative	0.825

In Error column 0.99 means there is 99% chance of that instance to test negative which is true and 1% possibility that the instances could test positive.

"+ "means prediction came out untrue. However, in the second instance 67% chance for the instance to test negative as compared to the instance in which we have 99% surety hence it proved wrong.

0.67 is not to close to 0.99 which gives the algorithm a benefit of doubt as to predict positive or negative.
Zero R

ZeroR is the simplest classification method. It is that type of classification method which would lean on the target and ignore other attributes.

TABLE V. PREDICTION TABLE

inst#	actual	predicted	Error prediction
1	1:tested_negative	1:tested_negative	0.648
2	2:tested_positive	1:tested_negative	+0.648
3	1:tested_negative	1:tested_negative	0.648
4	1:tested_negative	1:tested_negative	0.648
5	2:tested_positive	1:tested_negative	+0.648

We always get the same result for every instance either 65% (0.352 test negative) or 35% (0.352 test positive)

B. Logistic Regression

Logistic regression was developed by statistician David Cox in 1958. Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function. [7]

TABLE VI. PREDICTION TABLE

inst#	actual	predicted	error prediction
1	1:tested_negative	1:tested_negative	0.981
2	1:tested_ positive	1:tested_ positive	0.517
3	1:tested_negative	1:tested_ positive	+0.5
4	1:tested_negative	1:tested_negative	0.721
5	1:tested_ positive	2:tested_positive	0.582
6	1:tested_negative	1:tested_negative	0.841
7	1:tested_ positive	2:tested_positive	0.921
8	1:tested_negative	2:tested_ negative	0.927

C. Random Forest

Random forest generates many single classification trees. To classify a new object from an input, put the input vector down each of the trees in the forest. Each tree generates their own results and then they select one set of a class. [15]

```
plas< 111.5
|  preg< 7.5
|  |  skin < 29.5
|  |  |  age < 30.5
|  |  |  |  skin <19.5 :tested_negative (122/0)
|  |  |  |  skin >= 19.5
|  |  |  |  |  plas<94.5 :tested_negative (40/0)
|  |  |  |  |  plas>= 94.5
|  |  |  |  |  |  mass <32.7 :tested_negative (18/0)
|  |  |  |  |  |  mass >= 32.7
|  |  |  |  |  |  |  preg<0.5 :tested_positive (2/0)
|  |  |  |  |  |  |  preg>= 0.5
|  |  |  |  |  |  |  |  skin <23.5 :tested_positive (1/0)
|  |  |  |  |  |  |  |  skin >= 23.5 :tested_negative (10/0)
|  |  |  |  |  pedi>= 0.22
|  |  |  |  |  |  mass <37 :tested_positive (15/0)
|  |  |  |  |  |  mass >= 37
|  |  |  |  |  |  |  pres< 89
|  |  |  |  |  |  |  |  skin <36.5 :tested_negative (5/0)
|  |  |  |  |  |  |  |  skin >= 36.5 :tested_positive (2/0)
|  |  |  |  |  |  |  pres>= 89 :tested_positive (3/0)
|  |  |  plas>= 146.5 :tested_positive (48/0)

Size of the tree : 189
```

TABLE VII. PREDICTION TABLE

inst#	actual	predicted	Error prediction
1	1:tested_negative	1:tested_negative	0.58
2	1:tested_negative	1:tested_negative	0.55
3	1:tested_negative	1:tested_negative	0.95
4	1:tested_negative	1:tested_negative	1
5	1:tested_negative	1:tested_negative	0.6
6	1:tested_negative	1:tested_negative	0.81
7	1:tested_negative	2:tested_positive	+0.83
8	1:tested_negative	2:tested_positive	+0.65

D. Multilayer Perception

It works on how different attributes results process and interact with one another and alter their results in such a way that the final outcome is the filtered through each node (neuron).

Multi-Layer perception bestows great advantages as it is used for pattern classification, recognition, prediction and approximation.

TABLE VIII. PREDICTION TABLE

inst#	actual	predicted	error prediction
1	1:tested_negative	1:tested_negative	0.692
2	1:tested_negative	1:tested_negative	0.583
3	1:tested_negative	1:tested_negative	0.942
4	1:tested_negative	1:tested_negative	0.954
5	1:tested_negative	1:tested_negative	0.945
6	1:tested_negative	1:tested_negative	0.894
7	1:tested_negative	2:tested_positive	+80.85
8	1:tested_negative	2:tested_positive	+0.55

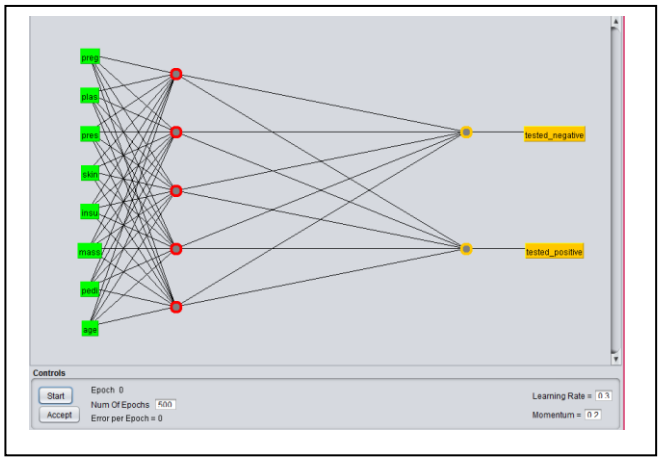


Fig. 2. Neural network (MLP)

Here we can see a network of different layers namely input layer, hidden layer and output layer consisting of input nodes

(green) or "neurons", output nodes (yellow) and some hidden nodes(red) some of them are visible. The nodes in the network are all sigmoid .Each connected network has some value in it which will be pass on to other nodes and each nodes perform a weighted sum of its input and pass it on until we get some results.Hidden layer depends upon the complexity of the data.[5].

Here MLP does show result with minimum error rate but it processes slow as compared to others.

E. J.48

Jr8 is basically an implementation of C4.5 algorithm [6] J48 decision tree decides which attributes is the most decisive one and which one is least and over and then these attributes further divided into sub tree. It generates a binary tree, unlike Random Forest decision tree. It use the concept of entropy, difference in entropy gives us the attribute which is free to make decisions.

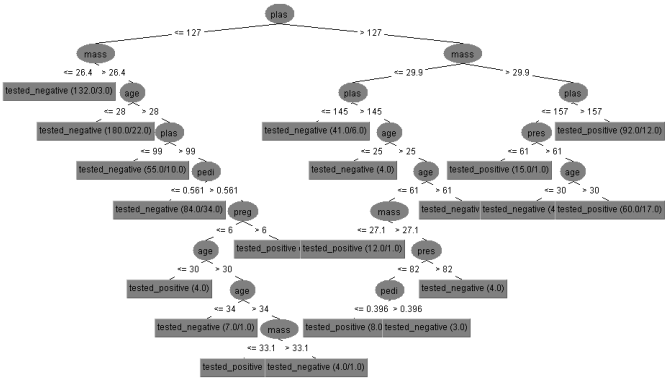


Fig. 3. Descion tree (j48)

TABLE IX. PREDICTION TABLE

inst#	actual	predicted	error prediction
1	1:tested_negative	1:tested_negative	0.982
2	2:tested_positive	2:tested_positive	0.635
3	1:tested_negative	2:tested_positive	+0.635
4	1:tested_negative	1:tested_negative	0.867
5	2:tested_positive	1:tested_negative	+0.9
6	1:tested_negative	1:tested_negative	0.867

VI. CLASSIFICATION RESULTS

Positive = identified

Negative = rejected.

Therefore:

TP=True positive = correctly identified

FP=False positive = incorrectly identified

TN=True negative = correctly rejected

FN=False negative = incorrectly rejected.[6]

Accuracy= (TP + TN) / (TP + FP + TN + FN) [8]

TABLE X. COMPARISON OF ACCURACY BETWEEN DIFFERENT ALGORITHMS

S.No	Classification Type	Accuracy (%)
1	Naïve Bayes	76.3 %
2	MLP	81.8182%
3	J.48	75.3%
4	ZeroR	67.5%
5.	Random Forest	79.2%
6.	Regression	76.8%
7.	Logistic Regression	79.2%

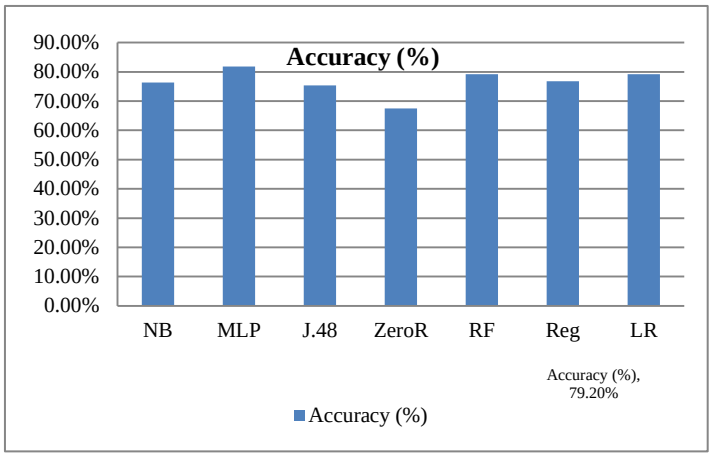


Fig. 4. Graphical representation of Accuracy over different algorithm.

VII. CONFUSION MATRIX

TABLE XI. COMPARISON OF CONFUSION MATRIX BETWEEN DIFFERENT ALGORITHMS

Algorithm	TN	FP	FN	TP
Naïve Bayes	45	7	10	15
MLP	49	3	11	14
J.48	43	9	10	15
ZeroR	52	0	25	0
Random Forest	45	7	9	16
Logistic Regression	45	7	9	16

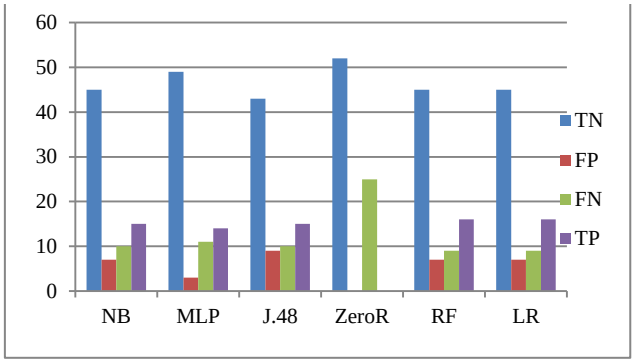


Fig. 5. Graphical representation of Confusion Matrix over different algorithm.

TABLE XII. COMPARISON OF ABSOLUTE ERROR BETWEEN DIFFERENT ALGORITHMS

S.No	Error
Naïve Bayes	0.2841
ZeroR	0.4481
J48	0.3036
Logistic	0.2867
Random Forest	0.3884
MLP	0.244

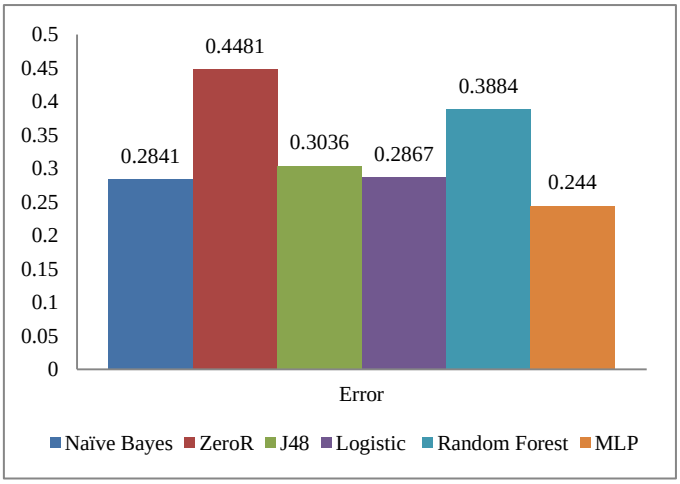


Fig. 6. Graphical representation of absolute error over different algorithm.

VIII. CONCLUSION AND FUTURE SCOPE

In order to make effective and efficient results, we need to work on a different algorithm and to make sure which suits best. Diagnosing diabetes through data mining tool over medical records of patients though it has been done by a majority of the researchers but the more we dig into the matter the furthest we can go.

In terms of performance we find multi layer perception function most effective hence it shows fewer errors however it takes too much processing time because it requires calculation of weights of each node. ZeroR is useful to determine baseline performance for others classification method. Naïve Bayes is also very efficient as it gives a predominant result after each validation but its performance is not quit impressive. J4.8 gives a graphical image of the precedence of the attribute as it calculates the priority of each attribute with other and yet it also predicts accurate results with least error hence it requires time.

Our objective of comparing the algorithm on the same dataset, analyzing and predicting the results out of it has been achieved.

In future, we will be interested in gathering information among our own neighborhood and we were keen to get new results which lead us toward more precise and accurate divination.

REFERENCES

[1] WorldHealth Organization, Diabetes Programm http://www.who.int/diabetes/en/

[2] Machine Learning Group at the University of Waikato..Weka 3: Data Mining Software in Java. Retrieved September 4, 2016 from http://www.cs.waikato.ac.nz/ml/weka/

[3] Sanofi, Diabetes Pakistan, Statistics http://www.sanofidiabetes.com.pk/web/about_diabetes/statistics

[4] The News International https://www.thenews.com.pk/print/73051-seven-million-pakistanis-suffering-from-type-2-diabetes

[5] Ian H. Witten, Department of Computer Science University of Waikato New Zealand, " Simple neural networks", "More Data Mining with Weka".More Data Mining with Weka, Simple Neuarl Network, https://drive.google.com/file/d/0B-f7ZbfsS9-xcEFUZ095UUpnVlU/edit

[6] Sathees Kumar B , Gayathri P,Department of Computer Science,Bishop Heber College ,"Analysis ofAdult-Onset Diabetes Using Data Mining Classification Algorithms", International Journal of Modern Computer Science(IJMCS)ISSN: 2320-7868 (Online)Volume No.-2, Issue No.-3, June, 2014Conference proceeding.

[7] Logistic Regression https://en.wikipedia.org/wiki/Logistic_regression

[8] P. Radha , Dr. B. Srinivasan, "Predicting Diabetes by cosequencing the various Data Mining Classification Techniques", IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 6, August 2014.

[9] PimaIndiansDiabetesDataSethttps://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes

[10] Aiswarya Iyer, S. Jeyalathaand Ronak Sumbaly "Diagnosisof Diabetes Using Classification Mining Techniques" , International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015

[11] N.Satyanandam,Dr.Ch. Satyanarayana,Md.Riyazuddin, Amjan.Shaik "Data Mining Machine Learning Approaches and Medical Diagnose Systems ", International Journal of Computer & Organization Trends – Volume2Issue3-2012

[12] Sadri Sa'di, Amanj Maleki, Ramin Hashemi, Zahra Panbechi and Kamal Chalabi, "Comparison Of Data Mining Algorithms In The Diagnosis Of Type II diabetes", International Journal on Computational Science & Applications (IJCSA) Vol.5, No.5,October 2015

[13] D. Ezaz Ahmed, Dr. Y.K. Mathur, Dr Varun Kumar, "Knowledge Discovery in Health Care Datasets Using Data MiningTools", (IJACSA) International Journal of Advanced Computer Science and Applications,Vol. 3, No.4, 2012117.

[14] Sukhjinder Singh, Kamaljit Kaur, "A Review on Diagnosis of Diabetes in Data Mining", International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438.

[15] Tahani Daghistani,Riyad Alshammari, "Diagnosis of Diabetes by Applying Data Mining Classification TechniquesComparison ofThree Data Mining Algorithms", IJACSA) International Journal of Advanced Computer Science and Applications,Vol. 7, No. 7, 2016.

Comparative study on sentiment analysis using WEKA classifiers

Rabica Shamim
Dept. Computer Science and Information Technology, NED University of Engineering and Technology
Karachi, Pakistan
Rabica_shamim@yahoo.com

Syeda Sadaf Shahab
Dept. Computer Science and Information Technology, NED University of Engineering and Technology
Karachi, Pakistan
shahabsadaf@hotmail.com

Sohail Abdul Sattar
Dept. Computer Science and Information Technology, NED University of Engineering and Technology
Karachi, Pakistan
sattar@neduet.edu.pk

Abstract*— Classification of text is an important aspect of sentiment analysis and helpful in making decision about any product, article, book, movie etc available in market. There are different algorithms available for classifying text data. We have tested a dataset comprising of 2000 text files containing movie reviews[1] using Neural Network (NN), Support Vector Machine (SVM), Naïve Bayes (NB), Logistic regression and J48 algorithm in supervised learning environment using WEKA on training sets and testing test to proposed the best suitable algorithm for text classification in positive and negative sentiment. In this paper we will define the relevancy of above mentioned algorithms, test results which we have performed, comparative study of different results we observed and conclusion.

Keywords*— text mining; algorithm for text mining; decision making;

I. INTRODUCTION

Contextual base analysis classifies the data with respect to polarity and the opinion given in text. It is widely used in online shopping cart usually for product reviews; provide a buyer to make their mind to purchase or not about particular item.

There are multiple classification algorithm, they are only useful if we use right algorithm for right category of data which provides which provides optimal and efficient results.

In next paragraph we will provide a brief description about the algorithm we have used to conclude best option for text mining.

A. Neural Network

Neural Networks is a widely accepted paradigm to analyze trends and pattern. The best part of NN is that it results are unbiased and doesn't worry the distribution of the data.

B. Support Vector Machine

SVMs are learning method introduced by V. Vapnik[3]. It is best suitable supervised machine learning algorithm which is suitable for classification of text based problem.

C. Naïve Bayes

This is based on Bayes' algorithm using conditional probability for analyzing each result of observation.

D. Logistic Regression

Logistic regression is best suitable for mine the binary data, its output is discrete.

E. J48

J48 is a branch from decision tree; it is a powerful technique in decision making and data extraction. It uses training data set or extracting information benefits and it examines the results from choosing an attribute for splitting the data. The splitting procedure stops if all instances in a subset belong to the same class. Then the leaf node is created in a decision tree telling to choose that class

II. CLASSIFICATION

WEKA classification classifies the data in to training set which allow choosing classifiers. For data classification we need to provide the following:

- Data set
- Division of data into training set
- Identify the suitable classier
- Understanding of training data result

III. EXPERIMENT

We have performed the experiment using test data to predict the optimal data mining algorithm using WEKA. It is working with different format; Data can be loaded from various sources, including files, URLs and databases. Supported file formats include WEKA‟s own ARFF format, CSV, Lib SVM‟s format, and C4.5‟s format. It is also possible to

generate data using an artificial data source and edit data manually using a dataset editor. We had data in text files which we had converted into .arff format using WEKA SimpleCLI using following commands:

- Open SimpleCLI from Weka GUI chooser
- write:
- java weka.core.converters.TextDirectoryLoader –dir Location_of_Text_Data_files_folder\ > Location_to_save_arff_format_data\Name_of_arff_file.arff
- press enter

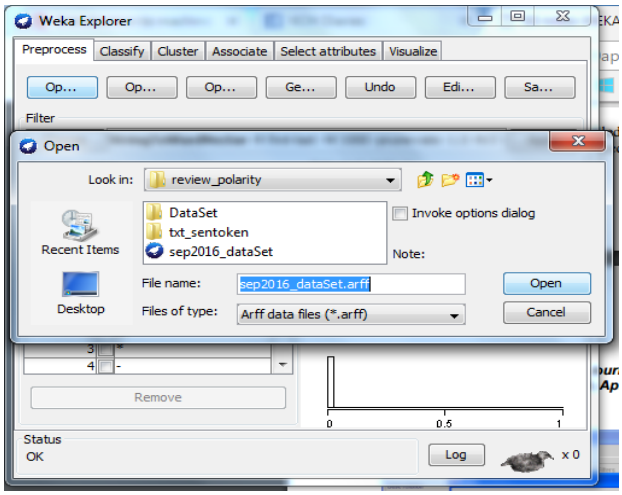


Figure 1: Loading Data File in Weka

This procedure generates the .arff files of data set provided. Load the .arff file using WEKA Explorer see figure 1.

Filter option allow to filter data as per requirement, here had chosen *StringtoWordVector* which divides the text into positive and negative classes showing in figure 2.

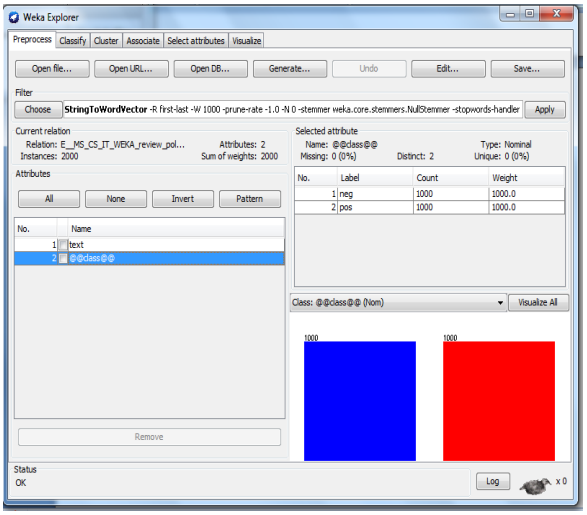


Figure 2: Filtered Data in Positive and Negative classes

IV. COMPARATIVE ANALYSIS

Java provides open source Opinion mining to analyze evolutionary algorithms for big data. Main issues in mining include techniques of data classification. To evaluate data we used two modes: 1. K fold and 2 Percentage spilt mode. In K fold database is randomly divided into K disjoint blocks of objects , then the algorithm is trained using k- 1 blocks and the remaining blocks is used to test the performance of the algorithm. This process is repeated k times. In last all records averaged. In percentage split, the database is randomly split in to two disjoint datasets. The first set, called training set where as second set is called test set; it is common to randomly split a data set under the mining task in to 2 parts. We divided the dataset into 75% training set & 25% test set. Once the tests is carried out using the selected datasets, then using the available classification and test modes ,results are collected and an overall comparison is conducted.

We performed the testing using multiple classifiers their correctly classification of instances vary. In our experiment we observed that J48 accuracy level for Correctly Classified Instances is lowest where as Naïve Bayes accuracy level was highest among five algorithms used.

Not only correct instance recognition Naïve Bayes has lowest Incorrectly Classified Instances count, Table 1.

=== Summary of Naïve Bayes ===

Correctly Classified Instances	1688	84.4%
Incorrectly Classified Instances	312	15.6 %
Kappa statistic	0.688	
Mean absolute error	0.1989	

Table 1: Comparative results of Algorithms

Coverage of cases (0.95 level)	98.1 %
Total Number of Instances	2000

Using Table 1 as reference we can nominate the best suitable algorithm for text polarity classification, i.e.

1. Naïve Bayes

Data Type (Test Data)	**Naive Bayes %**	**J48 %**	**Logistic R %**	**SVM %**	**NN%**
Accuracy	80.6	71.8	80.2	80	78.2
Non Accuracy	19.4	21.2	19.8	20	21.8

2. Logistic regression
3. SVM
4. Neural Network
5. J48

V. CONCLUSION

In this paper we analyzed text dataset to review the best optimal Text extraction tool. Weka helped to analyze the large text files using various mining algorithms. Test result suggests that '*Naïve Bayes'* is optimal solution for text based data with low rate of incorrect instance finding and high rate of correct one. We can use naïve bayes algorithms in online stores (for product review), statement analysis, Personality analysis questioner etc.

VI. REFERENCES

[1] http://www.cs.cornell.edu/People/pabo/movie-review-data/

[2] M. Govindarajan, and R. M. Chandrasekaran, "Classifier Based Text Mining for Neural Network", World Academy of Science, Engineering and Technology International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol:1, No:3, 2007

[3] Vladimir N. Vapnik. The Nature of Statistical Learning Theory. Springer, New York, 1995.

[4] Dr. Neeraj Bhargava, Girja Sharma, Dr. Ritu Bhargava and Mansih Mathuri, Decision Tree Analysis on J48 Algorithm for Data Mining", International Journal of Advanced Research inComputer Science and Software Engineering, Volume 3, Issue 6, June 2013.

[5] Trilok Chand Sharma and Manoj Jain, "WEKA Approach for Comparative Study of Classification Algorithm", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 4, April 2013

[6] Neelamadhab Padhy , Dr. Pragnyaban Mishra , and Rasmita Panigrahi, "The Survey of Data Mining Applications And Feature Scope", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012

[7] Swasti Singhal, and Monika Jena, "A Study on WEKA Tool for Data Preprocessing, Classification and Clustering", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-2, Issue-6, May 2013.

[8] Nur Hafieza Ismail, Fadhilah Ahmad , and Abdul Azwa, "Implementing WEKA as a Data Mining Tool to Analyze Students' Academic Performances Using Naïve Bayes Classifier", UniSZA Postgraduate Research Conference 2013, At UniSZA